35

University of Cape Town
Faculty of Engineering and the Built Environment
Civil Engineering Department

Application of Survival Analysis Modeling in Water Pipelines Failure in Cape Town

A thesis submitted to the Faculty of Engineering and the Built Environment
Department of Civil Engineering
University of Cape Town

In partial fulfillment for the requirements of the degree of Master of Philosophy (MPhil)

Salma. M. E. Abdelgadir

Thesis Supervisor: Professor Romano Del Mistro

February 2012

Acknowledgment

I am deeply thankful to all who supported me during the completion of this work.

My gratitude is expressed to my supervisor Professor Romano Del Mistro. He has kindly and generously provided me with high standard supervision during the course of my study; and all the support, confidence and encouragement I needed. His objective and constructive guidance has greatly improved my learning perspectives and scope.

Much appreciation is directed to the staff of the University of Cape Town for continuously seeking to pave the ways of knowledge for the researchers and to provide a loving environment for study.

I am also indebted to Mr Mamoun Elkheir and Mr Bahaadin Elhassan for their valuable statistical advice and help which is most appreciated.

Most importantly, I would like to express my gratefulness to every member in my family. Their support and encouragement have always uplifted my failing spirits. To my father Mr M. E. Abdelgadir, my husband Hassan and sons Mohamed, Ali and Omer; my brother Mamoun and sisters Sara and Yousra: your strong belief, continual prayers and patience have provided me with all the spirit that I needed.

I know the meaning of plagiarism and declare that all the work in this document, save for that which is properly acknowledged, is my own

Abstract

The statistical modelling of water pipelines failure has been widely adopted by water authorities in their pursuit to proactively manage their aging water distribution systems. In this thesis failures of the 100 mm FC pipes in Cape Town City have been modelled by using survival analysis techniques. Estimates of the Mean Cumulative Function (MCF) have been used to predict the failure rates of pipes in the network. The failure rate was found to follow the bathtub curve and three different regression models were derived for each stage of it and these are:

- The Regression equation for Age < = 25 years:

$$FR_1(t) = 0.008t - 0.028$$

- The Regression equation for 25 < Age < 60 years:

$$FR_2(t) = 0.417$$

- The Regression equation for Age > 60 years:

$$FR_3(t) = 0.003t^2 - 0.174t$$

The probability of survival of pipes has been predicted in relation to different factors that are assumed to have an influence on the failure of pipes by applying a Proportional Hazard Model (PHM) to the case study dataset. These factors include the pipe length (L), the installation eras of pipes and the number of previous failures (NOPF). Among these factors, the number of previous failures was found to have the most effect on the failure probability. Longer pipes were found to be more prone to failure while some constructions of manufacture defect are suggested to influence the failure of younger pipes. The predicted failure rates and hazard rates revealed that pipes in the network have a failure pattern that is similar to the Bathtub Curve.

Contents

List of figures

List of Tables

Dedication

To my Parents

Who taught me the meaning and pleasure of knowledge!
With their prayers I have reached every success in my life,

To my Supervisor

For his patience, persistent support and help; and for granting me the opportunity to satisfy my passion for knowledge;

To my Family

For their persistent support and encouragement,

Acknowledgment

I am deeply thankful to all who supported me during the completion of this work.

My gratitude is expressed to my supervisor Professor Romano Del Mistro. He has kindly and generously provided me with high standard supervision during the course of my study; and all the support, confidence and encouragement I needed. His objective and constructive guidance has greatly improved my learning perspectives and scope.

Much appreciation is directed to the staff of the University of Cape Town for continuously seeking to pave the ways of knowledge for the researchers and to provide a loving environment for study.

I am also indebted to Mr Mamoun Elkheir and Mr Bahaadin Elhassan for their valuable statistical advice and help which is most appreciated.

Most importantly, I would like to express my gratefulness to every member in my family. Their support and encouragement have always uplifted my failing spirits. To my father Mr M. E. Abdelgadir, my husband Hassan and sons Mohamed, Ali and Omer; my brother Mamoun and sisters Sara and Yousra: your strong belief, continual prayers and patience have provided me with all the spirit that I needed.

I know the meaning of plagiarism and declare that all the work in this document, save for that which is properly acknowledged, is my own

Abstract

The statistical modelling of water pipelines failure has been widely adopted by water authorities in their pursuit to proactively manage their aging water distribution systems. In this thesis failures of the 100 mm FC pipes in Cape Town City have been modelled by using survival analysis techniques. Estimates of the Mean Cumulative Function (MCF) have been used to predict the failure rates of pipes in the network. The failure rate was found to follow the bathtub curve and three different regression models were derived for each stage of it and these are:

- The Regression equation for Age < = 25 years:

$$FR_1(t) = 0.008t - 0.028$$

- The Regression equation for 25 < Age < 60 years:

$$FR_2(t) = 0.417$$

- The Regression equation for Age > 60 years:

$$FR_3(t) = 0.003t^2 - 0.174t$$

The probability of survival of pipes has been predicted in relation to different factors that are assumed to have an influence on the failure of pipes by applying a Proportional Hazard Model (PHM) to the case study dataset. These factors include the pipe length (L), the installation eras of pipes and the number of previous failures (NOPF). Among these factors, the number of previous failures was found to have the most effect on the failure probability. Longer pipes were found to be more prone to failure while some constructions of manufacture defect are suggested to influence the failure of younger pipes. The predicted failure rates and hazard rates revealed that pipes in the network have a failure pattern that is similar to the Bathtub Curve.

Contents

List of figures

List of Tables

1 Introduction

Background

Ageing water pipelines and their increasing costs have become worldwide problems. The challenge is how to manage buried water distribution systems in an effective, reliable and sustainable manner. Water pipeline failures can be associated with numerous hazards that affect economics, public health and the broad environment. Failure of pipes results in: loss of water; property damage; interruption of service; decreased system performance and the financial cost of restoring the failed pipe. It has, thus, become inevitable for water utilities to assess, maintain and upgrade the current conditions of their water supply systems.

For several decades, the prevailing manner to manage water distribution systems was reactive. Since the late 1970's and beginning of 1980s, concerns have been directed at the increasing number of breaks in water pipelines. Many studies have been done to investigate failure modes of pipes and their causes. Others aimed at modelling the failure rates and their subsequent economic implications. However, with the increasing demand for water supply as well as the extreme degradation of water network's reliability, water utilities have tended to adopt pipeline asset management practices that help them to develop long term management strategies for their water distribution networks. This would include, inter alia, the condition assessment of water networks, improvement in data collection and management, predictive modelling for failure, risk management and long term budget allocation for maintenance and rehabilitation requirements of the pipes.

Distribution networks often account for up to 80% of the total expenditure involved in water supply systems. As water mains deteriorate both structurally and functionally, their breakage rates increase, network hydraulic capacity decreases, and the water quality in the distribution system may decline (Kleiner & Rajani, 2001). According to the latest version of the Water Services Development Plan for Cape Town (WSDP), (COC, 2008), the replacement cost of the water reticulation system represent approximately about 70% of the replacement cost of the entire water supply infrastructure. About 15% of the capital budget for 2010 has been earmarked for water infrastructure and wastewater management. The City is systematically replacing ageing water pipes. It has been reported that over the past few years the City has managed to accelerate pipe replacements from 7.4 km per year to 33.8 km (as at April 2008), (COC, 2009). This only amounts to a 0.5% replacement of

the metropolis's water networks of 10500 km length. The international norm is between 1% and 2% of the total network.
This thesis aims to develop failure prediction models for Cape Town water pipelines models that help Cape Town's water authority to proactively manage its water pipelines system. The thesis also provides a review of most of the existing models for pipe failures in the literature with a brief discussion of their mathematical and prediction characteristics.

A general equation for the failure rate of pipes was required in order to describe failure trends of the pipelines in relation to their ages. Different approaches and methodologies of regression analysis that vary from simple regression passing through survival analysis methods have been applied in this work in order to obtain the appropriate model that describes water pipelines failure in Cape Town.

The effect of the different predicting variables on the distribution of the time to event (i.e. time to failure of the pipeline) was also investigated by applying a Proportional Hazard Model (PHM) to the data. The PHM can be used to predict the failure time for individual pipes in the network as well as for groups of pipes of certain properties. Such models can help water utilities pin point the main factors that leads to the failure of the water pipelines and to adopt the appropriate management strategies for maintenance, rehabilitation or replacement of pipes in the network.

2 Literature review

2.1 Introduction

This chapter reviews different aspects related to water pipeline failures. In sections 2.2 and 2.3, general information about pipeline distribution systems and their failures mechanisms and causes have been presented. This includes a description of the different parts of the distribution system and their functions, followed by an overview on water pipeline's failures. The deterioration process, the factors influences this deterioration, failure modes and consequences of failure have all been discussed. The methods and approaches that have been used worldwide for managing water pipelines systems have been reviewed in section 2.4. Processes of data collection and condition assessment for water networks have been illustrated and the use of these processes in developing pipeline decision supports systems has been discussed. This revealed the importance of the use of statistical models in managing the ongoing deterioration of water pipeline networks and to support the decision making process for rehabilitation and replacement of pipes. A wide range of different pipe failure models in the literature have been reviewed in section 2.5 in terms of their dependent and independent variables. Survival analysis modeling, the method of modeling in this thesis, has been reviewed separately in chapter 3.

2.2 Pipeline Distribution Systems

Distribution system infrastructure is generally considered to consist of pipes, pumps, valves, storage tanks, reservoirs, meters, fittings, and other hydraulic appurtenances that connect treatment plants or well supplies to consumers' taps. The systems of pipes that transport water from the source (such as a treatment plant) to the customer are often categorized from largest to smallest as transmission or trunk mains, distribution mains, service lines and premise plumbing. Transmission or trunk mains usually convey large amounts of water over long distances such as from a treatment facility to a storage tank within the distribution system. Distribution mains are typically smaller in diameter than the transmission mains and generally follow the city streets. Service lines carry water from the distribution main to the building or property being served. Service lines can be of any size depending on how much water is required to serve a particular customer and are sized so that the utility's design pressure is maintained at the customer's property for the desired flows. Premise plumbing refers to the piping within a building or home that distributes water to the point of use (NRC, 2006).

There have been several descriptions for the reliability of the pipe network presented in the literature. The general feature is that the system should fulfill the following requirements:

- The system should deliver the required quantity of water for both human purposes as well as emergency flows (e.g. firefighting) at an acceptable pressure. This requirement may be interpreted as the "hydraulic integrity" of the system.
- The water quality provided by the system should comply with the safety criteria specified by regulation and standards. This requirement may be interpreted as the "water quality integrity" of the system.
- The system should withstand all external and internal stresses acting upon it. This would imply the "physical integrity" of the system.
- The system should be economically efficient, (Kleiner Y. , 1997; Røstum, 2000; NRC, 2006).

The physical, hydraulic and water quality integrity of a system are influenced by the specific structural properties of pipes, the surrounding environment as well as the operational regimes. The structural properties of pipes may include pipe installation; material; diameter; length; wall thickness; hydraulic carrying capacity; bedding conditions and laying depth. On the other hand the environmental effects may include soil characteristics, pipe location, land use, climate, and water corrosivity. The operational regimes refer to the water pressure, back flow potential, flow velocity and maintenance and rehabilitation strategies and histories applied to the network (Røstum, 2000; NRC, 2003).

The main purpose of most of the extensive research works done in the last few decades throughout the world was to explore the relationship between the single and combined effect of these factors in water mains failures. The use of the statistical modelling of failure data can be considered as the most convenient and effective way. The physical investigation of every single segment of pipe is neither possible nor economically viable. These models can then be used by water utilities in making decisions for maintenance and rehabilitation of pipes in the long term; hence plans for the associated budgets can be prepared. An overview of the main issues related to pipe failure will be discussed in the following sections.

2.3 An Overview of Water Pipeline Failures

This section presents an overview of water pipeline failure. A general description of the deterioration of pipes is firstly reviewed. This is followed by an explanation of the major factors influencing pipelines failure and the different widely experienced failure modes and consequences.

2.3.1 Deterioration of Water Pipeline System

As in other engineering systems, the structural conditions of water pipeline systems deteriorate over the system's lifetime. This deterioration mainly affects the system in two ways. Firstly, it affects the pipe's structural resilience and its ability to withstand the various types of stresses imposing upon it. This will, at some certain time in the future, lead to the failure of some parts of the system. Secondly, it affects the inner surfaces of the pipes resulting in diminished hydraulic capacity, degradation of water quality and reduced structural resilience in cases of severe internal corrosion. Both categories of deterioration contribute to diminish the reliability of the distribution network, (Kleiner & Rajani, 2001). In water pipeline systems, deterioration is often evident from one of the following signs, (Collicot, 2004):

- Frequent breaks (by visual observation)
- High leakage rates (by visual observation)
- Reduced hydraulic capacity (e.g. low pressure complaint)
- Impaired water quality (e.g. rusty water, unfavorable odour or taste)

Examples of different types of deterioration in water pipelines are illustrated in Figure 2-1:

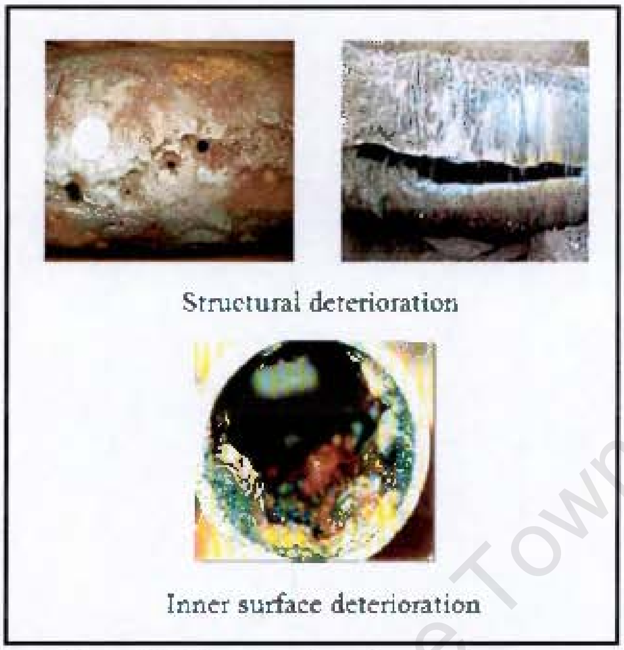


Figure 2-1: Deterioration of water pipelines.
Source: (Collicot, 2004)

The deterioration process of water mains has been described by (Misiunas, 2005) as consisting of several stages:

1. Installation;
2. Initiation of corrosion (in metallic pipes);
3. Crack before leak due to either corrosion or mechanical stresses;
4. Partial failure; and
5. Complete failure.

Misiunas, (2005) indicated that this sequence of failure might not be applicable to all pipes. Some pipes (e.g. steel and ductile iron) are likely to leak before they crack, while others typically break before they leak (e.g. cast iron, larger diameter pre-stressed concrete and PVC pipes). The deterioration of water pipes is often described in reliability engineering by the so-called Bathtub Curve (Wasson, 2006), which will be discussed further in section 2.5.2.3.

2.3.2 Factors Influencing Pipe Failure

It has been discussed in section2.2 that the integrity of water pipelines may be influenced by several factors which can generally be grouped under three categories as structural, environmental and operational factors. These factors are well known in the literature as playing a key role in the degradation, and eventually, in the failure of pipes. It is, hence, essential for water utilities to identify these factors and to incorporate them in the analysis of failure data. This will depend on each network's special features and characteristics and on the availability of the relevant data, (Alonsoa et al., 2008). In the literature, these factors are referred to as covariates or independent variables.

However, many researchers indicated the lack of deep understanding of the physical mechanisms and the interactions of various factors contributing to breaks. Some of these factors can be quantified and hence it is convenient to include them in the analysis, such as age, diameter, length and laying depth. Other factors include improper beddings, stresses from high internal pressure or changes of material properties overtime. Where applicable, these factors may be included in the analysis as dummy variables. Undoubtedly, incorporating all the factors in the analysis and investigating their effects in breaks of pipes may help pinpoint the main reasons of failures. Therefore, more accurate results are expected to be obtained and changes in pipeline designs and construction polices may be issued, (Shamir & Howard, 1979; Andreou, 1987; Kleiner, 1997).

In this section, factors that are thought to influence pipe failure will be reviewed, with a brief discussion about their effects as has been presented in the literature. Some of the potential interactions between some of the factors that lead to pipe failure are shown in Figure 2-2.

Loading

Water mains may encounter different types of loads, which can be categorized as external and internal. External loads may generate from soil pressure, frost pressure, traffic loads, stress due to temperature differences along the pipe and third parties while internal loads may be induced due to the operational fluid pressure (Yaminighaeshi, 2003). Stresses due to the movement of the ground or disturbance of the pipe bedding are other types of loads imposed upon pipes, (Boxall, O'Hagan, Pooladsaz, & Saul, 2007). Mains are designed to withstand all these anticipated external and internal forces. However, structural failure can occur if the actual forces exceed the structural strength of the pipe material due to poor design, deterioration of pipe strength, or unanticipated forces, (O'Day, 1982).

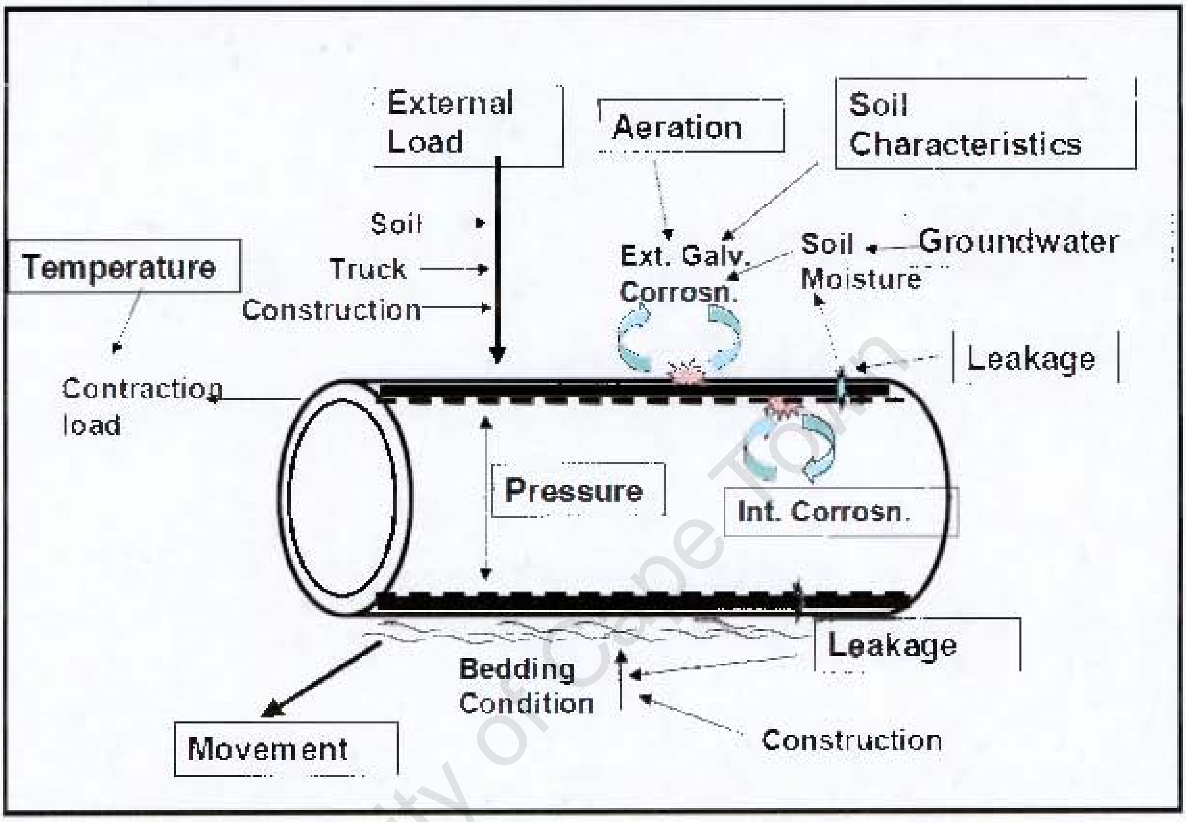


Figure 2-2: Potential interactions leading to failure of water mains.
Source: (O' Day, 1988)

Material

Materials that have been used in water pipelines can generally be categorized as metallic, cementitious and plastic. Examples of metallic pipes are steel, cast iron, ductile iron and copper. The common types of cementitious pipes used are asbestos cement, reinforced and pre-stressed concrete pipes. Significant proportion of pipes in many water utilities throughout the world are asbestos cement (AC) and cast iron (CI) pipes (grey/ductile). Plastic pipes have only being used since the 1970s. Representatives of plastic pipes are Polyvinyl Chloride (PVC) and Polyethylene (PE), (Kleiner, 1997; Røstum, 2000; McGrath & Mruk, 2002; NRC, 2003; NRC, 2006).

Material has widely been considered as one of the most important leading factors of the failure of water pipelines. This is due to the role that material properties play in resisting internal and external stresses that a pipe may experience during its service life (Boxall, O'Hagan, Pooladsaz, & Saul, 2007).

The increasing number of failures in cast iron and asbestos cement pipes has motivated many researchers to monitor their failure patterns, (e.g. Kettler & Goulter, (1985); Boxall, (2007); DeSilva et al. (2003)). This has been joined with the fact that these are the oldest and most frequently used pipes in many networks in the past. Nowadays, the use of plastic pipes is more common. This might be attributed to benefits such as corrosion resistance, improved hydraulics, life expectancy, durability and reduced installation costs (Uni-Bell, 2003). Other factors that have been mentioned in the literature related to material are corrosion, wall thickness, and manufacturing techniques.

In Cape Town, the vast majority of pipes are Fibre Cement (FC), which represents around (72%) of the network. The next major materials are Cast Iron (CI), and Cast Iron Concrete Lined (CICL) which represent around (7.8) % of the network. The remaining 20% of the network include materials such as Steel (ST), UPVC, HDPE and others

It has been reported that fibre-cement pipes manufactured before 1985 are more prone to bursts due to the type of materials used in the manufacturing process. Pipes manufactured after 1985 were modified to prolong service and reduce failures. However much of the City's fibre-cement pipes manufactured prior to 1985 is failing, although a significant number are still functioning efficiently. According to the Utility Services Portfolio Committee Chairperson, Cllr Clive Justus, a blanket replacement of fibre-cement pipes will not be financially prudent (COC, 2007).

Age

Since the very beginnings of water mains breakage analysis, age has taken the largest share of attention as a key factor affecting failure (e.g. Shamir & Howard, (1979) and Kettler & Goulter, (1985)). Pipes may encounter different chemical, operational, and environmental conditions during their service life. The ability of pipes to withstand all these conditions imposed them may be compromised as the pipes age (Hu & Hubble, 2007). However, controversy about the significance of age as a major cause of failure was shown among other researchers. It has been reported in Walski & Pelliccia, (1982) and O'Day, (1982) that age was considered as a minor factor in some previous studies in the United States.

Age may also be used as an indicator of the installation period, manufacture and workmanship quality which are also being believed to affect pipe failure significantly.

Diameter

The relationship between break rate and diameter of pipes has been reported in many studies. Clark, (1982) and Kettler & Goulter, (1985) concluded that higher break rates are associated with smaller pipes diameter. This might be explained by the structural characteristics of smaller diameter pipes such as reduced pipe strength, reduced wall thickness, reduced joint reliability, lower velocities and lower resistance to chemical attacks, (Kettler & Goulter, 1985; Wengström, 1993; Røstum, 2000; Hu & Hubble, 2007). In contrast, in the analysis of failure data of the two water utilities made by (Andreou, Marks, & Clark, 1987), pipe size was not found to be related to break rate for one system, while in the second system larger diameter pipes were found to experience a faster third break. These findings were attributed to the worst overall condition of the second system. Generally, large and small diameter pipes are experiencing different failure modes due to the different loading conditions associated with each size. These modes were discussed in details in (O'Day, 1982).

Length

The contribution of pipe length to the failure of pipes has not been clearly concluded in the literature. In most reported analyses of water mains failure, it has been used as a normalizing factor (break/year/unit length), (e.g. Shamir & Howard, (1979); Walski & Pelliccia, (1982); Clark, (1982). This would lead to the expectation that the number of breaks is directly proportional to the length of pipes. In other applications, length has been used as covariate. Andreou, (1987), used the natural log of the length as a covariate in his proportional hazard model and found that the hazard rate was approximately proportional to the square root of length. The paper indicated the following regarding the relationship between length and pipe breakages:

- Length might, up to a certain degree, be a surrogate for land development activities that were not reflected in the land development variables provided in the data sets. That is, longer lengths might be associated with sections of the pipe being in more remote areas of the system with less land development (and thus less stress-causing factors). This point was also mentioned by Skipworth, (2002), where the source referred to main length as a scale of the degree of urbanization and to reflect the spatial density of potential weak points in the network (e.g. increased concentration of service connections, pipe junctions, valves, joints etc.). If these connections are considered as weak points, then shorter pipes may indicate a higher break rates.

- On the other hand, Skipworth, (2002) also indicated some localized forces that affect breakage of pipes. Such forces could be caused for example by non-uniformity of the bedding and differential soil subsidence, restraints by bends, branches and valves, thermal and moisture changes in the bedding material and tree root growth pressure. Thus, a pipe break could indicate a localized concentration of those forces in the vicinity of the break and not necessarily the presence of high risk factors along the whole pipe length. Therefore, when break causing factors are localized, it shouldn't be expected that the predicted break rate to vary linearly with the length of the pipe.

Corrosion

Corrosion in buried pipes may occur as a result of the different interactions between pipe material, soil components and the conveyed water components. Corrosion of cast iron mains has been defined by O'Day, (1982) as the electrochemical reaction (graphitization) between the pipe metal and its environment; the pipe loses its ferrite constituent, leaving behind the graphite. In asbestos cement pipes corrosion is caused by a number of chemical agents which attack the inner and outer sides of pipes walls. These agents include acids, sulphate, magnesium salts, alkaline hydroxide, ammonia and soft water, (Nesbar, 1983; Jarvis, 1998).

In general, (Srikanth.S, 2005) Srikanth.S, (2005), identified the following six types of corrosion that can occur in a buried pipeline: (1) pitting corrosion because of material in-homogeneities, (2) chloride or sulphate induced stress corrosion cracking, (3) corrosion by concentration cells in soil arising out of differences in oxygen concentration in the soil adjacent to the pipe at different regions, (4) microbiologically induced corrosion under anaerobic conditions by sulphate-reducing bacteria (SRB) and acid-producing bacteria (APB), (5) tuberculation because of the buildup of corrosion products on the internal pipe surfaces and (6) stray current corrosion by earth return direct currents.

Corrosion is well known as a major cause of failure in water mains. It reduces the structural strength of pipes by reducing the thicknesses of their walls and over time a pipe will not be able to resist all the forces acting upon it. However, different methods of lining have been extensively used to protect pipes against corrosion. The effect of corrosion on pipe failure and models that have been developed in this regard is discussed in details by Yaminighaeshi, (2003).

Installation conditions

Improper installation is considered as one of the major reasons of the failure of PVC pipes. Excessive glue inside the pipe, over or short insertion and wrong antifreeze are the common causes that lead to failure (Priddy, 2008). Poor bedding conditions may cause beam failure in small diameter pipes. This may happen due to different circumstances (e.g. poor initial construction, erosion of the bedding by joint leakage, soil movement from the shrink-swell of expansive soils or nearby excavations) (O'Day, 1982).

Soil conditions

Soil characteristics were found to impact pipes in terms of corrosion. NRC (2003) reported results from surveys conducted in the United States and Canada to determine the most common causes of external corrosion of water mains. It was found in that survey that 67% of the respondents perceived that corrosive soils are the primary cause of external corrosion of water distribution mains in their systems. Highly corrosive soil was also pointed to have an impact on higher break rate for all pipes in the study undertaken by Andreou, (1987).

Climate

Studies of water main breaks histories for many cities indicate that a drop in seasonal temperature is almost always followed by an increase in the number of breaks. The ways that lower temperatures in winter months affect mains are: (Walski & Pelliccia, 1982; O'Day, 1982; Hu & Hubble, 2007):

- Increased earth load caused by frost
- Increased tensile stress on mains caused by temperature-induced contraction
- Increased external stresses caused by soil moisture expansion from frost penetration

- Rainfall as a surrogate of soil moisture which may eventually result in erosion of bedding

However, (Røstum, 2000) indicated that climatic effects should be used at a preliminary stage in order to determine pipe failure causes since it is not easy to include it in the analysis as a covariate.

Number of pervious failures (NOPF)

One of the most important indicators of pipe failure is the number of previous failures. That is, once a pipe experiences a break it is more likely to break again. This fact was first reported by (Walski & Pelliccia, 1982) and subsequently by others (e.g. Clark (1982); Andreou (1987); Goulter & Kazemi (1988)). This might be attributed to similarity in condition for pipes in the same location. It is expected for pipes in the same location to have the same age and materials, to be laid with the same construction and joining methods and they are also likely to be exposed to the same external and internal corrosion conditions (Røstum, 2000). A report on water mains break proposed by NRC, (2003) included the information shown in Table 2-1.

Table 2-1: Potential information in a water main failure report.
Source: (NRC, 2003)

General	Location	Physical Data	Type of Failure	Probable Cause of Failure	Type of Repair
• Date and time break reported • Time when water was shut off • Time when water was turned on • Properties affected • Air temperature • Repair by • Property damage • Broken fitting	• Nearest property address • Distance from nearest property line • Distance from nearest intersection • Northing and easting • Isolation valves operated	• Pipe diameter • Pipe material • Year of installation • Pipe wall thickness or pipe class • Type of lining • Type of joint • Type of water service • Normal operating pressure • Under boulevard or road • Depth of cover • Depth of frost • Type of native soil • Type of backfill • Soil resistivity • Soil sample collected (Yes / No) • Pipe sample collected (Yes / No)	• Circumferential break • Longitudinal break • Split bell • Corrosion pit hole • Leaking joint • Leaking valve • Leaking service connection	• Corrosion • Ground frost • Joint failure • Disturbance (third party) • High pressure • Frozen pipe	• Repair clamp • Replace pipe section • Replace valve • Replace service connection • Anode installed • Repair joint

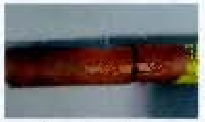
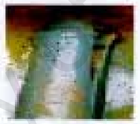
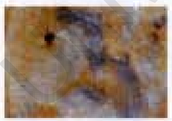
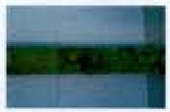
2.3.3 Failure Modes

The failure modes of water mains have been described repeatedly in the literature. The common types of failure may be categorized as follows (Rajani, Zhan, & Kuraoka, 1996; NRC, 2003):

1. Circumferential failures;
2. Longitudinal failures;
3. Through hole failure;
4. Rapture blowout failure;
5. Bell sheer failure;
6. Spiral failure;
7. Joint failure; and
8. Corporation rock failure.

These failure modes occur as a result of the different internal and external stresses acting on pipes. Rajani et al. (1996), identified these stresses as direct tension failure, bending or flexural failure and hoop stress failure. Other failure may befall as a result of corrosion or any third party reasons (e.g. traffic accidents). Table 2-2 shows the different failure modes and the combined responsible stresses.

Table 2-2: Typical failure modes of pipes

Failure mode	Example photographs of pipes	Possible causes stresses
Circumferential break	 Cast iron pipe  Copper pipe	• Direct tension stresses due to: ▪ Soil/traffic/frost loads ▪ Frictional resistance ▪ Soil shrinkage ▪ Thermal expansion or contraction ▪ Pitting due to corrosion • Bending stresses due to: ▪ Soil/traffic/frost loads ▪ Poor bedding ▪ Poor compaction ▪ soil movement
Longitudinal break	 Cast iron pipe  Asbestos cement pipe  PVC pipe	• Hoop stress due to: ▪ Internal water pressure ▪ External freezing water pressure • Ring stress due to: ▪ Soil/traffic/frost loads ▪ Increased ring loads from expansion of frozen moisture in the soil
Through hole break	 Cast iron pipe  Ductile iron pipe  PVC pipe  Copper pipe	• Corrosion/ Graphitization pits in metallic pipes • Internal pressure/chemical attack in plastic pipes

Continued: Table 2-2.

Sources: (Rajani et al., 1996; Rajani & Kleiner, 2001; NRC, 2003; Rajani & Kleiner, 2003; WSAA, 2003; Workman, 2009)

Failure mode	Example photographs of pipes	Possible causes stresses
Rupture blowout break	 Ductile iron pipe  PVC pipe	• Hoop stress due to: ▪ internal water pressure ▪ freezing of water pipe line ▪ Corrosion of steel pre-stressing in concrete pre-stressed pipes
Bell shear break	 PVC pipe	• Bending stress • Over homing spigot during construction
Joint failure	 PVC pipe	• Expansion of joint material • Freezing of water pipe line
Spiral break	 PVC pipe	• Bending stress • Hoop stress due to internal water pressure or freezing of pipes

Circumferential breaks are more likely to occur in smaller diameter pipes owing to the poor conditions of bedding or erosion of bedding due to water leaks. This type of failure is unlikely to happen in big diameter pipes which are more likely to experience ring or hoop stresses (O'Day, 1982). Failure modes of PVC pipes are most often catastrophic. In this type of material failure tends to cause extensive splitting along the full length of the pipe (Zamojc, 2005).

2.3.4 Consequences of Pipe Failure

Water mains breaks have many and varied consequences which affect both water utilities and customers. Water utilities may concern about the health, environmental and economic effects of pipe breaks, while customer's concerns may be directed at the quality and level of service. Losses of water due to the failure of pipes are of major concern for all water utilities, given the renewed worries about water scarcity throughout the world. This has been associated with the potential risk of pollution as a result of the conflux of contaminants and of corrosion of pipe walls and the reduction in fire fighting capability (Makar & Kleiner, 2000). From an economic point of view, pipe breaks are costly. Breaks require for immediate actions from municipalities to either repair or replace the broken segment. If this action has not been taken within a certain time, damages may extend to the adjacent properties and constructions which would lead to an increasing economic lost.

Other economic costs may happen due to the disruption of traffic flow if the break occurs in main roads or highways. From customer perspectives, breaks lead to disruption of service, traffic delays, less hydraulic capacity and potential hazard to the public health, (Shamir & Howard, 1979; Walski & Pelliccia, 1982; Makar & Kleiner, 2000; NRC, 2006).

In order to mitigate the different impacts of water pipeline failures, comprehensive strategies for managing water distribution system should be developed by water utilities. The operation and maintenance of the network must be governed by long term strategies that take into consideration the reliability of the system, customer's satisfaction and the budgetary constraints within the utility.

Included among these strategies would be the application of principles of municipal infrastructure asset management (Ingenium, 2006) which include basing decisions on an appropriate standard of service that can be afforded and level of risk that can be accepted;

supported by improved understanding of the probability of failure through modelling and non-destructive testing of pipes in the network.

2.4 Managing Water Pipeline Systems: Asset Management Approach

2.4.1 Introduction

It seems that there is a widespread agreement between water distribution system owners about the need for an efficient and effective management system for their networks, (Herz, 1998; Kleiner et al., 1998a,1998b; Hadzilacos et al., 2000; Herz & Kroop, 2002; Burn, et al., 2003; Moglia et al., 2006; Kleiner et al., 2007; Kleiner & Rajani, 2010). This is evident from the increasing concerns about the application of the Asset Management approach for water distribution systems. This has been demonstrated by the development of a number of decision support systems to help utility owners to manage their systems. Such asset management strategies should assist the network owners to evaluate the condition of the water distribution network, identify failure modes and risk of failure, identify areas of high risk, propose "repair or replace" strategies and prioritize the work based on the intrinsic risk and available budgets. Figure 2-3 shows a schematic diagram of the pipe failure management cycle as proposed by (Misiunas, 2005). A similar diagram describing the pipeline management cycle was made earlier by (Makar & Kleiner, 2000).

The application of Asset Management in water distribution systems would lead to the overall upgrade of the systems, allowing for the full utilization of pipes, keeping the system in good condition while addressing backlogs and minimizing the total lifecycle costs. The application of Asset Management strategy require for the following processes (NAMS, 2002):

- Collecting data (e.g. physical data and operation history);
- Identifying the desired levels of service and future demands;
- Assessing of the whole condition of pipes by using the available techniques;
- Identifying failure modes and assessing risks;
- Assessing the financial capabilities;
- Preparing the Asset Management Plans; and

- Reviewing and monitoring the whole process.

In the following subsections some of the Asset Management processes will be highlighted as applied to water distribution systems.

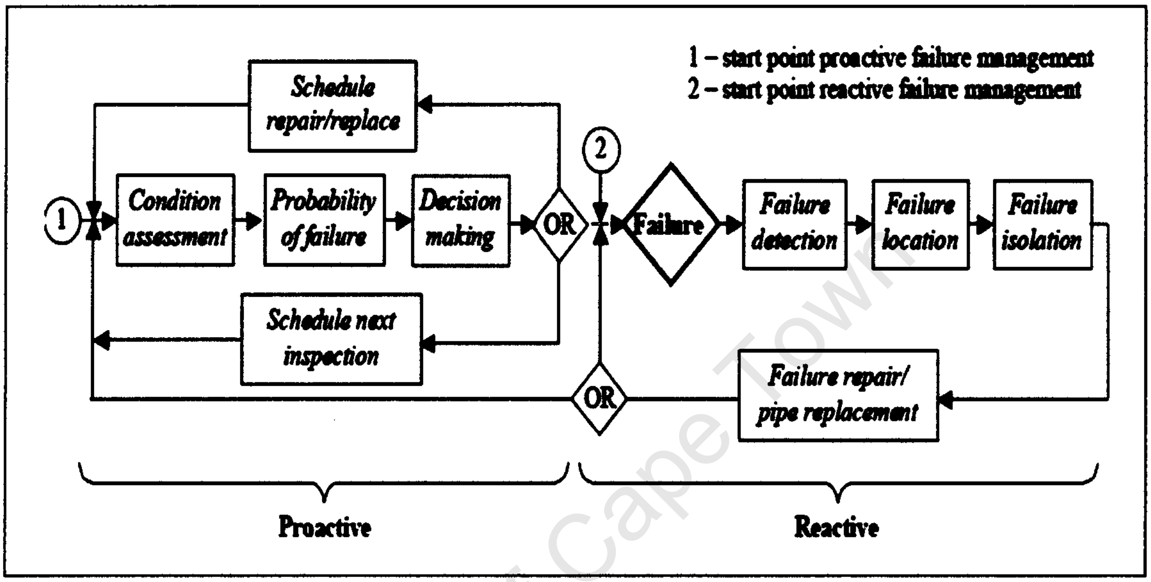


Figure 2-3: Pipe Failure Management Cycle.
Source: (Misiunas, 2005)

2.4.2 Data Collection

Data collection is essential for water utilities at any stage of the Asset Management process. Good quality data inventory would lead to more efficient and effective management strategies. In the context of this work, good quality data is needed for more accurate prediction of failure patterns. This includes the physical information about pipes in the system (e.g. location, diameter, material, etc.), operational data (e.g. pressure, hydraulic capacity), and maintenance records (e.g. time/location/type of failure, type of repair). The availability of such data would extend the capability and predictability of the developed models.

2.4.3 Condition Assessment of Pipes

It is widely accepted that the condition assessment of buried pipes is a costly and difficult task. This may be attributed to the fact that all pipes are under the ground and not easy to be individually subjected to visual inspection. Generally, the condition of buried pipes may be investigated through different phases. A preliminary assessment of water main breaks, customer complaints, unaccounted for water, and data on routine sampling and inspection which should be conducted each year, will identify both trends and the need for more detailed investigations (NRC, 2003).

The second phase involves a more detailed investigation of network specific problems such as reduced hydraulic capacity, reported defects in water quality due to corrosion or increasing failure rate. This in turn can be undertaken by the adaption of two approaches. The first approach is the direct inspection of pipes using "Non Destructive Evaluation". The second is the modelling techniques which can also be classified into two subgroups. The first is the "physical/mechanical" modelling of the pipe network which aims to improve the understanding of the structural performance of water mains. This mechanistic approach account for the external and internal loads acting upon the pipe (e.g. soil; frost; traffic; pressure) as well as other multiplying factors such as corrosion and temperature-induced stresses.

The other modelling technique is the statistical modelling which is usually used to explain, quantify and predict pipe breakage or structural failures (probability of failure) of water mains (Makar & Kleiner, 2000; Rajani & Kleiner, 2001).

Both, the physical and statistical modelling are based on the use of specific water distribution system data. The statistical modelling of pipeline failures constitute the main focus of the work presented in this thesis will be discussed in detail in section2.5.

From an economic point of view, Non Destructive Evaluation is time consuming and costly. Different pipe materials; soil characteristics; and construction methods may require different inspection techniques. However, for some critical pipes where the consequences of failure are catastrophic and the history of failure may not be available, visual inspection may represent the most economically attractive and inevitable approach (Makar & Kleiner, 2000; DeSilva et al., 2003; Eiswirth et al., 2001). Misiunas, (2005), identified two types of methods for Non Destructive Evaluation techniques namely, visual and non-visual methods; while DeSilva et al., (2003), sorted these techniques according to the material of pipes. A detailed description of these techniques is beyond the scope of this work; however, some of the common methods for both metallic and cementitious pipes can be illustrated as follows:

Metallic pipes

- Remote Field Eddy Current Tools: Used to determine the remaining wall thickness of cast iron pipes
- Magnetic Flux Leakage: Provides a means to rapidly assess pipe walls for fractures and pits by using a permanent magnet
- Radiometric Probes: Used for the detection of active water outflow due to significant changes in the surrounding soil moisture and for the assessment of cavities behind the pipeline
- Current Probe: Used for the detection of active water outflow. (Makar & Kleiner, 2000; Eiswirth et al., 2001; DeSilva et al., 2003).

Cementitious pipes

- Acoustic Emission Monitoring: Uses a pair or an array of hydrophones to detect for the sound of wire breaking in pre-stressed concrete pipes, which can then be used for estimating the number and rate for breaking wires. These hydrophones have the ability to detect and differentiate between other sounds such as traffic noise, water flow and construction noises (Makar & Kleiner, 2000).
- Ground Penetration Radar: A technique used to image both the wall thickness and to provide information about embedment conditions such as cavitation and compaction level in asbestos cement pipes, (DeSilva et al., 2003).

2.4.4 Rehabilitation of Water Mains

Water distribution system can be considered as a repairable system. A repairable system is a system that, when a failure occurs, can be restored to an operating condition by some repair process other than replacement of the entire system. If a pipe fails, it can either be replaced or repaired to restore the system to an operating state. A non-repairable system is one which is discarded and/or replaced after its first failure (Basu & Rigdon, 2000). The action of restoring the pipe condition after failure is called rehabilitation.The rehabilitation of water pipeline systems has been defined by (Røstum, 2000) as "*all methods for restoring or upgrading the performance of an existing pipeline system. This includes maintenance and repair as well as renovation and replacement of pipes*". Rehabilitation strategies can significantly affect the useful service life of the network. That is better rehabilitation strategies leads to longer service life of pipes.

Rehabilitation strategies may either be reactive or proactive in nature. Reactive strategies imply that pipes are rehabilitated after the failure occurs, while proactive strategies require long term planning for rehabilitation. The benefit of the reactive approach is that a pipe section realizes its full economic life whereas its disadvantages may include all the unplanned costs associated with failure. The development of failure prediction models would promote the adoption of proactive rehabilitation strategies, which would, ultimately support the provision of efficient, sustainable, reliable, affordable and competitive water supply services. There have been several methods used for the rehabilitation of water mains. In this section methods that have been used in the replacement and rehabilitation of pipes will be illustrated.

Replacement methods

Replacement methods are used to assist in reducing leakage as well as providing other benefits such as increased hydraulic capacity, clean safe water supply conditions and after replacement the pipe can be restored to a condition similar to a new pipe. Replacement methods can generally be divided into the following two groups, open trench and trenchless technology. (Røstum, 2000; Thornton et al., 2008):

- **Open trench technology**: In this case pipes are replaced by digging in the ground, discarding the old one and laying in the new one. This method is often extremely costly and in some cases completely impractical, as in the case of dense urban cities.
- **Trenchless technology**: In this case pipe replacement can be undertaken using "no dig" or trenchless technologies, which are usually cheaper and almost always less disruptive. Some of the methods of the trenchless technologies are discussed below:
  - **Slip Lining**: this is probably one of the simplest of no dig replacement techniques. In this case the old pipe is cleaned out and a new smaller diameter pipe is drawn through or pushed through the old one. The new pipe is of a smaller diameter and usually made of polyethylene (PE).

- **Pipe Cracking or Pipe Bursting**: this technology used to increase the hydraulic carrying capacity. The old pipe is prepared and then a conical wedge is drawn through ahead of the new pipe. In this way it is possible to use the old pipe as a guide for the new pipe; however the new pipe is actually larger than the old one.

Renovation methods

Renovation methods are used where the original fabric of a pipeline is incorporated and its current structure is improved. Two sub-groups of rehabilitation techniques exist, namely structural and non-structural methods. The tow sub-groups of rehabilitation are described below: (Røstum, 2000; Thornton et al., 2008)

- **Structural methods**: These methods improve the strength of the pipe and the resulting pipe can be considered as a new pipe. Methods of this group include:
  - Cured in place: in this technique a fabric tube is impregnated with a thermosetting resin before insertion into the host pipe. The resin is then cured in the host pipe to produce a rigid pipe within the host pipe;
  - Close-fit pipes: is another type of slip lining which involves inserting a thermoplastic tube that has been temporarily deformed to allow sufficient clearance for insertion into the host pipe. Other structural methods include continuous pipes and inserted hoses.

- **Nonstructural methods**: These methods do not significantly improve the strength of the pipe and the resulting pipe can be considered to remain as it was prior to the rehabilitation with respect to the structural condition. The functional performance is of course improved as a result of reduced hydraulic friction and improved water quality due to a new and smoother internal surface. This method includes:
- **Pipe cleaning methods**: e.g. air scouring; rotating chains, rods, and scraper trowels; and pigging.
- **Spray relining**: these methods imply the lining of pipes with cement mortar or epoxy.

2.4.5 Pipelines Decision Support Systems

Identifying criteria of which pipes to rehabilitate and when to rehabilitate them is a major concern for water utilities. Extensive research work has been done in the last three decades

Gustafson & Clancy, (1999), defined a break as "one or more failures requiring a single excavation to repair. Failures may be a leak, a cracked joint, a blowout, a split pipe, a circular crack, or multiples thereof. If they are repaired using one excavation, it is counted as one break".

2.5.2.2 Regression Analysis

Regression analysis is a popular statistical tool for the investigation of relationships between variables. Basically, it involves the use of mathematical functions to model and investigate the dependency of one variable called the ***dependent variable*** on one or more other observable variables, known as the *explanatory or* ***independent variables***. The purpose of regression is to investigate behavior and interactions between variables in order to search for a cause-and-effect relationship (Nathabandu & Rosso, 2008).

Regression modelling has been widely applied to water distribution systems in order to inspect the relationship between different pipe failure measures and failure contributing factors.

Two basic types of regression models are typically used in the various applications of statistics: linear and non-linear regression. The subcategories of these types may be classified as:

- **Linear regression**: this type can either be single or multiple which can be described as:
  - **Single linear regression:** in this case the dependent variable E(Y) is a function of only one independent variable (covariate) X and their relationship is linear. This type of regression takes the mathematical form:

    $$E(Y) = \alpha + \beta X \qquad \text{Equation: 2.1}$$

    The two constants, intercept and slope, are unknown and are to be estimated from a sample of Y values with their associated values of X, (Soong, 2004).

  - **Multiple linear regressions:** the dependent variable in this type of regression is a function of two or more independent variables. Two examples of the potential mathematical formulae of this type are:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \qquad \text{Equation: 2.2}$$

$$E(Y) = \beta_0 + {}^{\beta_1}/_{x_1} + \beta_2 {x_2}^2 \qquad \text{Equation: 2.3}$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ are regression coefficients (parameters) to be estimated from a sample of Y values with their associated values of $(x_1, x_2, \ldots, x_m)$, (Soong, 2004; Graybill & Iyer, 1994)

- Non-linear regression: the same concept is applicable to non-linear regression and hence it may also be a single or multiple regressions.

  an example of the mathematical formula for the single non-linear regression is as follows:

$$E(Y) = \beta_0 + \beta_1 e^{\beta_2} \qquad \text{Equation: 2.4}$$

where β_1, β_2 are regression parameters to be estimated from a sample of Y values with their associated values of x, (Graybill & Iyer, 1994)
an example of the mathematical formula for the multiple non-linear regression is as follows:

$$E(Y) = \beta_0 + \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 x_2} \qquad \text{Equation: 2.5}$$

Graybill & Iyer, (1994), noted that linear regression means the regression function is simultaneously linear in the unknown parameters β_i, and non-linear regression means the regression function is not simultaneously linear in the unknown parameters β_i. This note explained the structure of the second example of the multiple linear regressions.

The following are some measures of pipe failure. They are commonly referred to as Independent Variables (IV).

- Break Rate (BR): this function can be obtained either for individual pipes or a set of pipes by normalizing the number of breaks for pipe length and time (e.g. breaks/kilometer/year).
- Mean Time between Failures (MTBF): It only applies when the underlying distribution has a constant failure rate. The MTBF is the reciprocal of the break rate (Wilkins, 2002b), that is:

$$MTBF = 1/BR \qquad \text{Equation: 2.6}$$

- Rate Of Occurrences Of Failure (ROCOF): the ROCOF is the time derivative of the expected cumulative number of failures and is defined as:

$$v(t) = \frac{dV(t)}{dt} \qquad \text{Equation: 2.7}$$

Where $V(t)$ denotes the mean number of failures in the interval $(0, t]$. it follows that the ROCOF may be regarded as the mean number of failure per time unit at time t (Røstum, 2000).

Mean Time to Failure (MTTF): denotes the probability that a pipe will remain functioning (survive) in the time interval $(\mathbf{0}, \boldsymbol{t}]$or the probability that a pipe will not fail in the time interval$(\mathbf{0}, \boldsymbol{t}]$. It is also called the survival function. The regression method used to model this function is called survival regression modeling or regression modeling of time to event. The survival function is usually derived for non-repairable systems. Since water distribution can be considered as a repairable system, the system is considered as a new system after repair and hence a new survival function can be calculated (Park et al., 2008; Park et al., 2010). This assumption allows for the modeling of the time elapsed until the first failure occurs and the time to the second and subsequent failures. Special attention will be paid to this type of regression modeling, since this is the method which will be used in this study to model water pipeline failures in Cape Town metropolis area. Regression modeling of survival data will be discussed in details in chapter 3.

2.5.2.3 The Bathtub Curve

The concept of the Bathtub Curve has been introduced in reliability engineering to describe the service life profiles of human made systems, of which water distribution systems is an example. (Wasson, 2006), reported, according to some previous studies, that the origin of the Bathtub Curve dates back to the 1940s and 1950s in the embryonic stages of reliability engineering. It basically represents a plot of failure rates or hazard rates over the active service life of the equipment. The name was derived from the characteristic Bathtub Curve hazard rate profile illustrated in Figure 2-4. It may be applied to a single component in a system; a set of components; or an entire system (Røstum, 2000; Wilkins, 2002a).

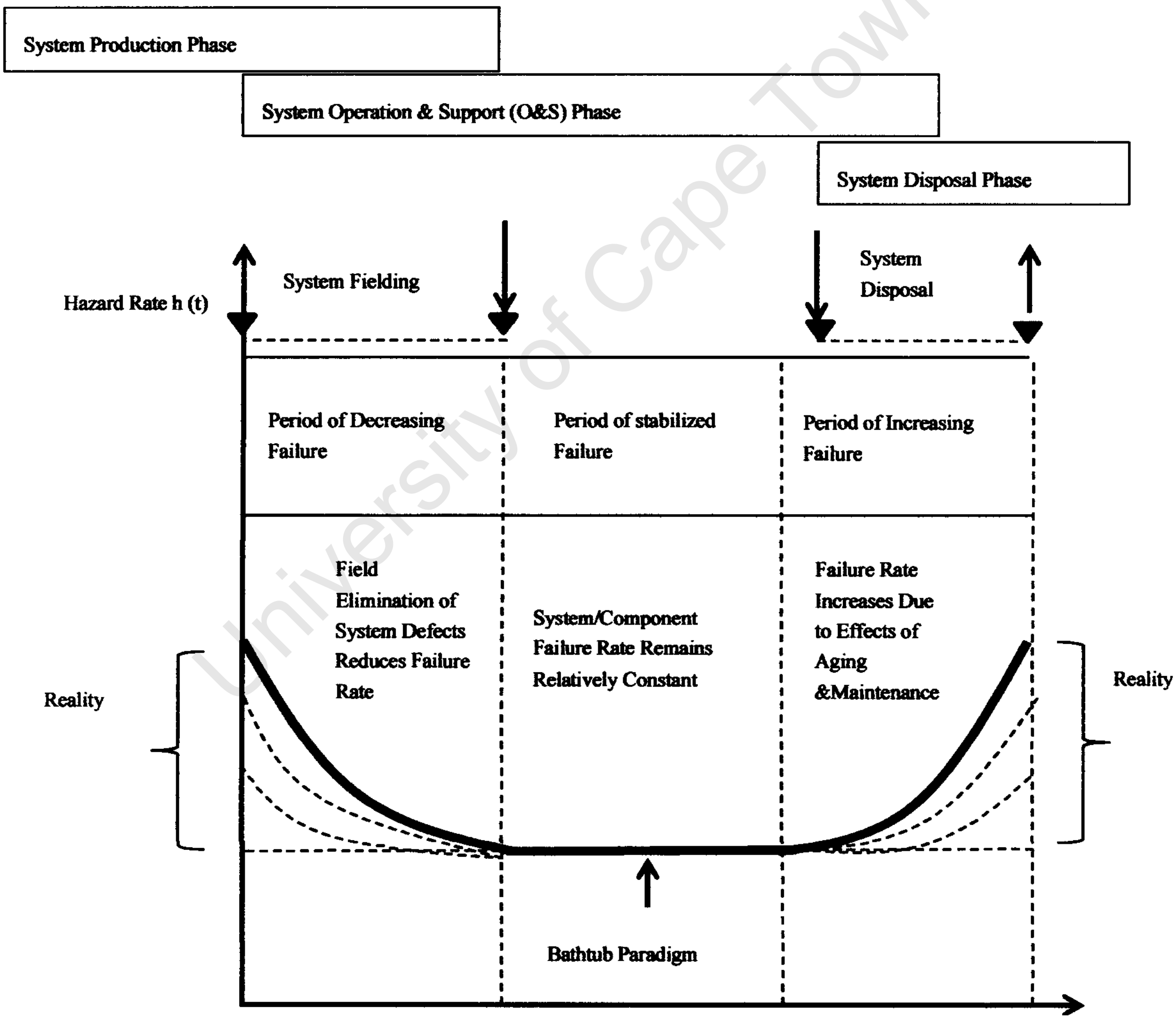


Figure 2-4: The Bath Tub Curve
Source: (Wasson, 2006)

Theoretically, the Bathtub Curve consists of three distinct service life phases:

- Period of decreasing hazard/failure rate: this period represents the period at the beginning of the service life of the system or the product; or the system fielding. Failures in this period usually occur as a result of manufacturing or installation defects. Therefore, this period is also known as the "Burn In" period; the "Infant Mortality" or the "Early Failure Period," (Wasson, 2006). Wilkins, (2002a), stated that the infant mortality does not necessarily mean that an object will fail during a certain specific time. It rather means that it is the period where failure rates decrease due to product control and defects elimination which may last for years.
- Period of stabilized hazard/failure rate: This period represents the normal operation and support of the system. It is also sometimes referred to as the "Useful Service Life" or the "Normal Life" period. This period is characterized by a low, relatively constant failure rate with failures that are considered to be random cases of "stress exceeding strength." Hazard rates in this period is therefore slightly stabilized, (Wasson, 2006).
- Period of increasing hazard/failure rate: This period is characterized by increasing failure rates due to the natural ageing and deterioration of components as well as some other third party factors, (Wasson, 2006).

There have been several discussions in the literature regarding the Bathtub Curve as a model of thinking in general; and as applied to water distribution systems specifically. Many authors criticized the idea that the Bathtub Curve is a one size mind-set that it applies universally to all systems and components. Some of the key issues were: (Røstum, 2000; Kleiner & Rajani, 2001; Wilkins, 2002a; Wasson, 2006):

- The Bathtub curve is typically used as a visual model to illustrate the three key periods of pipe failure and not calibrated to draw a graph of the expected behavior for a particular network. It is rare to have enough short-term and long-term failure information to actually model a network with a calibrated bathtub curve. Besides, the fact that water distribution system is a repairable system means that some pipes in the network might be rehabilitated or replaced; which results in discontinuity in the service life profile. Therefore, some of the existing models described the three stages of the bathtub curve while others described only one or two stages.

- There exists a great variance in failure trends between and within different networks and pipes due to the differences in structural; operational; and environmental conditions. That means that not every pipe in the network may experience the three stages of the bath tub curve and the length of each period may vary dramatically from pipe to pipe or from one network to another.

The shape of the bathtub curve may be slightly different, in the three stages, due to the effect of the different structural; operational; and environmental factors (referred to as reality factors in Figure 2-4. Putting the bathtub curve in mind may help municipal managers pinpoint the appropriate management strategies for water distribution systems in terms of product design and operational regimes. For instance a good operation and maintenance strategies may help the "In Service" period to be fully utilized and the "Wear Out" period to be postponed, while good product management may help the "Infant Mortality" period to be avoided or reduced, (see Figure 2-5).

2.5.2.4 Statistical Classification of Existing Pipe Failure Models

It has been indicated in sections 2.3.1 that the statistical modelling of pipe failure is one of the two subcategories of modelling techniques used in the condition assessment of the water distribution system and then in making decisions for the optimum management strategy. There have been different classifications and interpretations of the statistical techniques used in water distribution systems analysis. The majority of researchers classified the statistical models used for water distribution systems failures into two big groups: descriptive and predictive. Descriptive models are usually performed in order to identify the system characteristics; breakage patterns of pipes; and possible causal factors of failure. Every effort to model the structural conditions of the network should start with this basic kind of analysis. Predictive models may incur the same descriptive approach at its preliminary stages, as well as the prediction of possible future failures of pipe (i.e. probability of failure) (Andreou, 1987; Røstum, 2000; Pelletier et al., 2003; Stone et al, 2007). Predictive models have been categorized by Andreou, (1987); into three major categories:

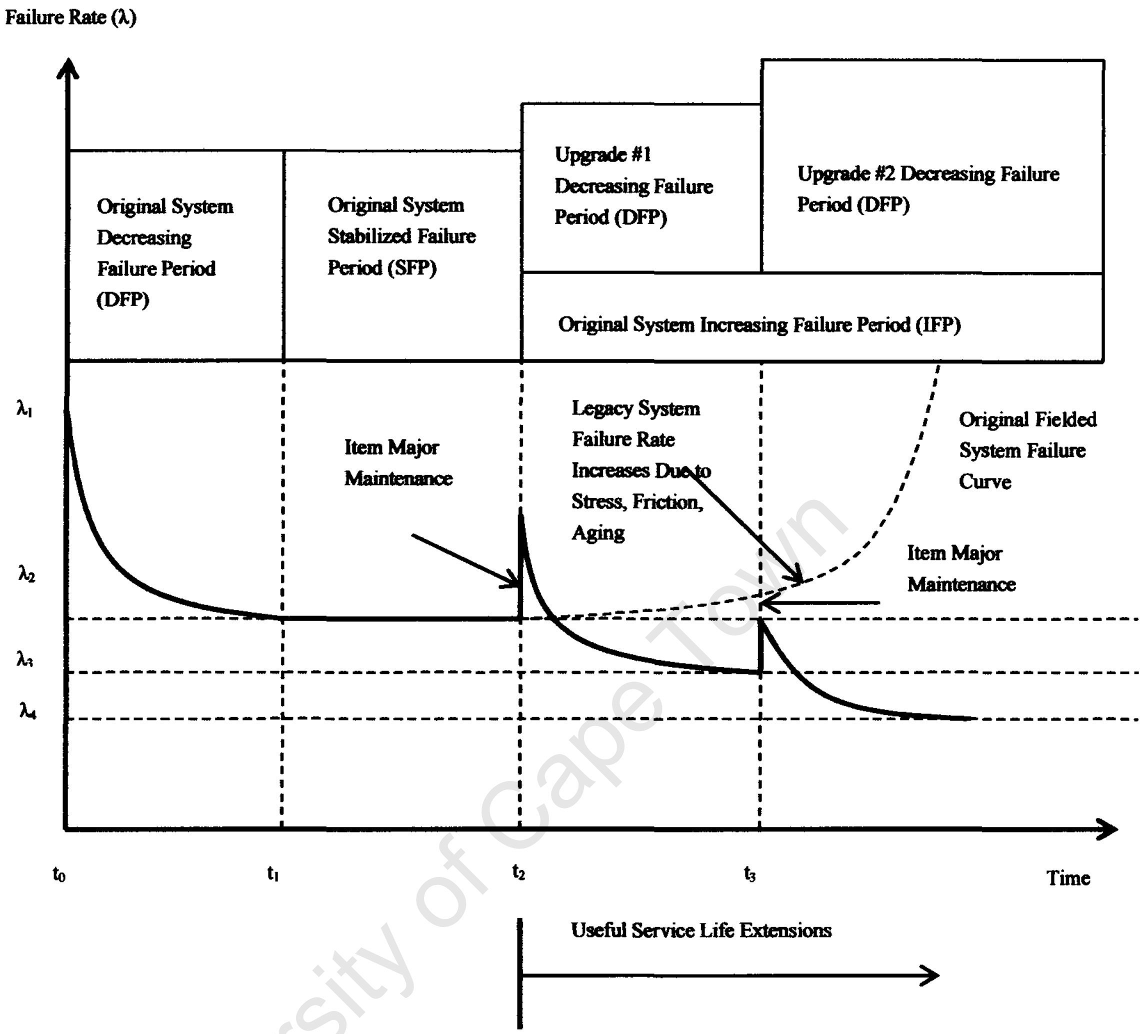


Figure 2-5: Service life extension
Source: (Wasson, 2006)

- **Aggregate models**: These are regression models where the expected number of breaks of a pipe (i.e. break rate) is a function of time t (i.e. age of pipe) since a reference time period and certain constant model parameters. Representatives of this type of modeling are the models developed by Shamir & Howard, (1979), where two regression equations were proposed (one exponential and one linear) to describe the break rate as a function of age. Although this type of modeling is characterized by its simplicity, the absence of the explanatory factors makes it difficult to develop insight about the break causing mechanisms and the key factors contributing to pipe failure (Andreou, Marks, & Clark, 1987). In terms of the classification of regression modeling discussed in section2.5.2.4, Shamir & Howard's models fall under the simple regression category.

- **Regression models**: These are also regression models where the expected number of breaks of a pipe (i.e. break rate) or the expected time to the next break (i.e. time between failures), are predicted as a function of certain independent variables reflecting environmental conditions and pipe characteristics. Representative of this type of modeling are the two regression models developed by Clark, (1982). A model was firstly proposed to predict the time elapsed from the installation of the pipe to the first failure. The other model was proposed to predict the expected number of breaks after the first failure, (Andreou, Marks, & Clark, 1987).
- **Probabilistic models**: This group include models that uses regression survival analysis to estimate the probability of failure/survival and then to assign the failure/survival time of pipes against a group of explanatory variables. Representative of this group of models is the model developed by Andreou, (1987) who identified two stages for deterioration and two models were developed to describe these stages. The first stage represents the early and slow stage of deterioration where the break rate was considered to be of less importance than the probability of a new break in any given time. A proportional hazard model was used to model the failure probability in this stage. The second stage represents the later and last stage of deterioration, where a Poisson-type model was found to fit the data best.

In terms of the classification of regression modelling discussed in section2.5.2.4, both the regression and the probabilistic pipe failure models fall under the multiple regression category.

The first two types of the predictive statistical modelling may also be referred to as deterministic models, that is, models developed by these methods can be used to determine the optimum management option on the basis of a determined (i.e. not probable) breakage pattern or breakage time, (Stone et al., 2007). "*In probabilistic models there is a certain amount of error in predicting one variable that cannot be explained in terms of other variables. Even if all conceivable variables are incorporated and any instrumental, observational, and recording errors are eliminated, there will still be some modeling error. This may be attributed to the incomplete knowledge of the physical processes involved or to some factors beyond control*" (Nathabandu & Rosso, 2008).

2.5.3 A Review of Some of the Existing Aggregate and Regression Models

Since the very first attempts for modeling water pipeline failure, (see Shamir & Howard, 1979; Walski & Pelliccia, 1982; Clark, 1982; and others), this area of research has received great attention from researchers in different places throughout the world. Significant number of models has been developed for several water distribution networks using different methodologies, different statistical procedures and models. A detailed review of these models is beyond the scope of this thesis as such a review has been extensively covered in the literature (see for example Røstum, 2000; Kleiner & Rajani, 2001;and Rajani & Kleiner, 2001b). However, a brief review of the work done by (Kleiner & Rajani, 2001) will be presented in this section; followed by a summary table of some of the models reviewed by them.

Kleiner & Rajani, (2001), provided a critical review of the statistical models that have been proposed in the literature to explain failure patterns, and to quantify and predict breakage rates and/or probability of failure in water mains. In this paper models that have been developed for water pipes failure were classified into two broad categories based on their dependent variables. Within each category there were subcategories which have been grouped according to the equation type of the model. The first category includes the deterministic models which have been grouped into time exponential and time linear models. The second category includes probabilistic models with subgroups of probabilistic single variate models and probabilistic multivariate models. A summary of these groups and subgroups is provided in Table 2-3 and a description of models of the first category (deterministic models) is presented in Table 2-4 and Table 2-5. Theory behind some of the models under the second category (probabilistic models) will be discussed in chapter3 and a review of some of these models will be presented in section 4.3.8.

Physically based models have been developed in order to improve the understanding of the physical/mechanical performance and behaviour of water mains. More details about this type of modeling can be found in the companion paper (Rajani & Kleiner, 2001).

Table 2-3: Categories of pipe failure models.
Adapted from: (Kleiner & Rajani, 2001)

Deterministic regression models	Probabilistic regression models	
Time exponential models	Probabilistic multivariate models	Proportional Hazard Model Accelerated Lifetime Model Time-dependent Poisson Model
Time linear models	Probabilistic single variate models	Cohort Survival Model Bayesian Diagnostic Model Semi-Markov Process Break Clustering Model

2.6 Conclusion

This chapter reviewed water distribution systems, their failures causes, mechanisms and consequences. General information about the pipeline distribution system has firstly been introduced. This is followed by an overview of pipeline failure. The deterioration process, factors influencing the failure, failure modes and consequences have been discussed. Several factors have been found to contribute in the failure of water mains. Such factors include pipe age, pipe material, pipe diameter, pipe length, soil condition, number of previous failures, etc. The effects of water pipelines failure on water utilities and customers have also been discussed.

Methods that have been used in managing failures of water distribution systems were reviewed with a special emphasis on the asset management approach. These include methods of data collection, condition assessment of pipes, rehabilitation of water mains and the development of decision support systems. The different statistical models and their associated methodologies used to help in the management and decision making process have been recited. Concepts such as the failure, regression analysis, the Bath Tub Curve have been presented and discussed.

Table 2-4: Some deterministic time exponential models.
Source: (Kleiner & Rajani, 2001)

References	Model	Notation	Data requirements
Shamir & Howard, (1979)	$N(t) = N(t_0)e^{A(e+g)}$	t = time elapsed (from present) in years $N(t) =$ No. breaks per unit length per year ($km^{-1}yr^{-1}$) $N(t_0) = N(t)$ at the year of installation of the pipe g = age of the pipe at the present time A = coefficient of breakage rate growth (yr^{-1})	Pipe length, installation date and breakage history; formation of homogenous groups essential according to criteria like pipe type, diameter, soil type, break type, overburden characteristics, etc.
Walski & Pelliccia, (1982)	$N(t) = C_1C_2N_{t_0}e^{A(e+g)}$	C_1 = ratio between {break frequency for (pit/sand spun) cast iron with (no/one or more) previous breaks} and {overall break frequency for (pit/sand spun) cast iron} C_2 = ratio between {break frequency for pit cast pipes 500 mm diameter} and {overall break frequency for pit cast pipes}	Same data as for Shamir and Howard (1979) plus information on the method of pipe casting and pipe diameter
Clark et al., (1982)	$NY = x_1 + x_2D + x_3P + x_4I + x_5RES + x_6LH + x_7T$ $REP = y_1e^{y_2t}e^{y_3T}e^{y_4PRD}e^{y_5DEV}SL^{y_6}SH^{y_7}$	x_iy_i = regression parameters, NY = number of years from installation to first repair D = diameter of pipe P = absolute pressure within a pipe I = % of pipe overlain by industrial development RES = % of pipe overlain by residential development LH = length of pipe in highly corrosive soil T = pipe type (1= metallic, 0 = reinforced concrete) REP = number of repairs PRD = pressure differential t = age of pipe from first break DEV = % of pipe length in law and moderately corrosive soil SL = surface area of pipe in low corrosivity soil SH = surface area of pipe in highly corrosive soil	Time of installation, breakage history, type and diameter of the pipe, as well as information about operating pressures, soil corrosivity and zoning composition of area overlaying pipe. Additional types of data such as the type of breaks and pipe vintage required to enhance model.

Table 2-5: Some deterministic time linear models.
Source: (Kleiner & Rajani, 2001)

References	Model	Notation	Data requirements
Kettler & Goulter (1985)	$N = k_0 \, Age$	N = number of breaks per year k_0 = regression parameter	Same data as for Shamir and Howard (1979)
McMullen (1982)	$Age = 65.78 + 0.028\,SR - 6.33\,pH - 0.049 r_d$	Age = age of pipe at first break (years) SR = saturated soil resistivity (Ω cm) pH = soil pH r_d = redox potential (millivolts)	Data required typically not available, sporadic data collection not expensive, however, continuous and extensive data collection program is costly; continuous monitoring of soil properties is important where ground water conditions have not reached steady state or are seasonally dependent
Jacobs & Karney (1994)	$P = a_0 + a_1 Length + a_2 Age$	P = reciprocal of the probability of a day with no breaks a_0, a_1, a_2 = regression coefficients	Pipe length, age and breakage history, more data enables formation of homogenous groups

3 Survival Analysis

This section presents a detailed description of Survival Analysis modeling in general, and as applied to the failure of water pipelines. General features and basic concepts of survival analysis will be first reviewed. This will be followed by an illustration of the different types of survival models and some examples of its applications.

3.1 What is Survival Analysis

Survival analysis has been defined by (Kleinbaum & Klein, 2005) as "*A collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.*" Time-to-event expression may be described as years, months, weeks, or days from the beginning of an observation of an individual until an event occurs. The event may be death or disease incidence in medical fields; failure of commercial banks in economics; or failures of pipes in water distribution systems. In survival analysis the time variable is usually referred to as the **survival time**, and the event is often referred to as failure. Equivalent expressions to the survival time are the **mean lifetime** or **life length**. However, in some medical applications the event may be a positive incidence, for instance remission from a disease after treatment or time to return to work after an elective surgical procedure. In this case, the time-to-event may be named the **response time**. (Kleinbaum & Klein, 2005; Borovkova, 2002; Lee & Wang, 2003)

The main interest when performing survival analysis is to characterize the distribution of 'time to event' for a given population (e.g. pipes), comparing this 'time to event' among different groups, or modeling the relationship of "time to event" to other covariates (prognostic factors or predictors), (Borovkova, 2002).

The statistical problem can be e characterized as either a recurrent event or a competing risk problem. A recurrent survival analysis considers outcome event(s) that may occur more than once over time for a given subject and the event may be caused by several factors. Examples of this type include transmission repairs in cars, recurrence of tumor bladder, and successive failures of water pipelines due to corrosion and traffic load. A competing risk survival analysis considers survival data in which each subject can experience only one of several different types of events (Nelson, 2003 ; Kleinbaum & Klein, 2005).

In section3.2), the basic statistical concepts used in survival analysis will be presented.

3.2 Basic Concepts in Survival Analysis

3.2.1 Time Origin

Emphases in the literature have been paid to three main requirements in order to perform survival analysis, (Cox & Oakes, Analysis of Survival Data, 1984; Borovkova, 2002; Machin, Cheung, & Parmar, 2006; Kleinbaum & Klein, 2005; Kalbfleisch & Prentice, 2002); namely:

1. A time origin must be clearly defined
2. A scale for measuring the passage of time must be agreed, and
3. The meaning of failure must be entirely clear.

The end point of the interval of the analysis may be easy to define unlike defining the time origin which may sometimes be a more difficult task (Machin, Cheung, & Parmar, 2006).

For instance, the time origin of water pipes to enter a survival analysis study may be when the first pipe was laid or at any other certain point in the lifetime of the water distribution system. The identification of a time origin for a specific network can be, to a large extent, more dependent on the availability of failure data. Many water utilities have just started their records on pipe failure in the last three decades, while the time origin of their networks may dates back to the early 1900s (Røstum, 2000).

3.2.2 Censoring, a Special Feature of Survival Analysis

There are some special features of survival data which make it amenable to standard statistical procedures used in data analysis. The first feature is that survival data are not symmetrically distributed. The survival distribution often tends to be positively skewed (i.e. the histogram will have a longer tail to the right of the interval that contains the largest number of observations). This feature can be attributed to the fact that the survival time of some individuals is unknown or **censored**. Censoring can be considered as one of the main features of survival analysis. The survival time of an individual is said to be censored when the end-point of interest has not been observed for that individual, (Collett, 2003; Kalbfleisch & Prentice, 2002).

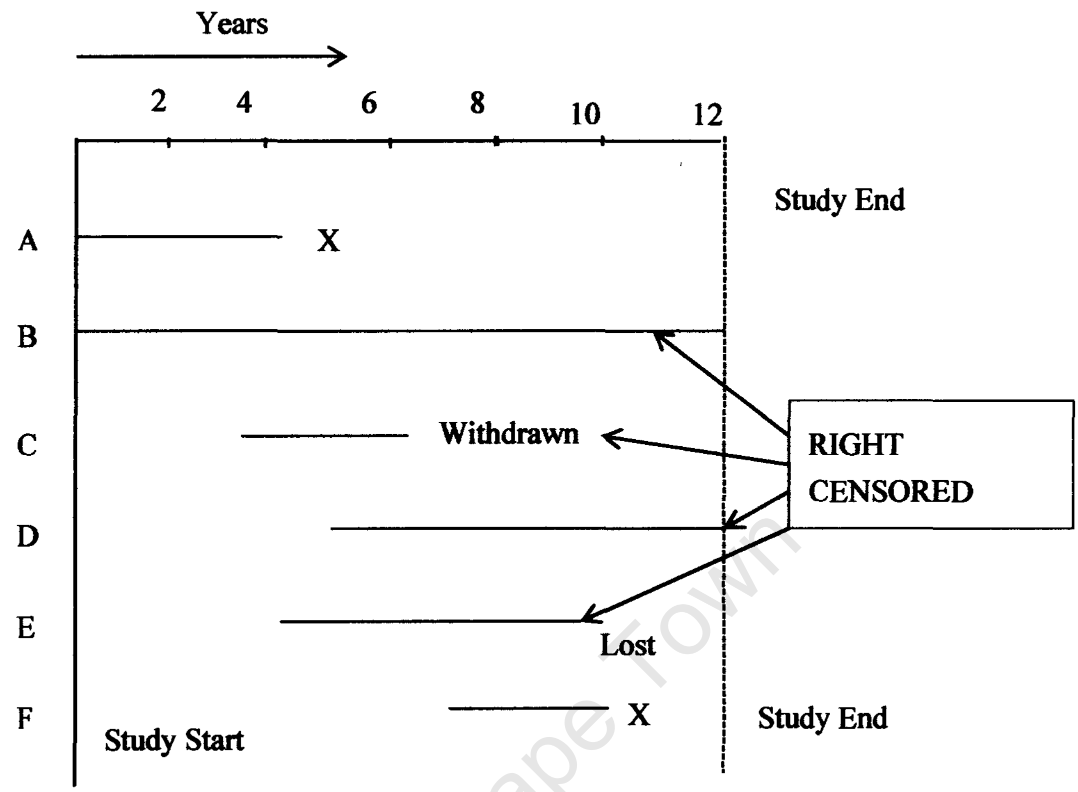
There are generally three reasons why censoring may occur, (Collett, 2003; Kleinbaum & Klein, 2005):

1. An individual does not experience the event before the study ends (i.e. a pipe hasn't failed yet);
2. An individual is lost during the study period (e.g. a pipe has been replaced by another type of pipe);
3. An individual withdraws from the study for any considerable reason (e.g. the pipe was found to have failed because of poor installation quality rather than corrosion in a study that aims to model pipe failure vs. corrosion).

In each of the three previous cases, censored survival time, although unknown, is observed after an individual entered a study; that is, to the right of the last known survival time, and therefore known as ***right censoring***, (Collett, 2003). Figure 3-1 illustrates the concepts of right censoring.

Assume that the individuals in Figure 3-1 are pipes and that the time origin of the study is the year of installation of pipes A and B. Pipe A has experienced a failure after 5 years of installation, which represent its actual observed survival time. Pipes B and D have not experienced any failures until the end of the study; therefore they are known to be ***right censored***. Pipes C, D, E, and F entered the study at different dates after the start of the study. In statistical terms this is known as ***staggered entry***, (Røstum, 2000). Pipe C has been withdrawn from the study for some reason while pipe E has been lost. The survival time of pipes C and E can be assumed to be the last day of survey, but the case for these two pipes can still be considered as right censoring (Collett, 2003), . Pipe F has experienced a failure after installation and before the end of the study so the actual survival time of this pipe can be determined and there is no censoring in this case.

Another two forms of censoring are ***left censoring*** and ***interval censoring***. ***Left censoring*** is encountered when the actual survival time of an individual is less than that observed (e.g. a pipe failed far before the failure is observed and recorded (see Figure 3-2). In ***interval censoring*** individuals are known to have experienced an event within an interval of time, (Collett, 2003). To illustrate the concept of interval censoring, consider a pipe, which is known to be repaired after a failure incidence, and another failure occurred and observed after three months. The actual time when that pipe failed is not determined but it is known that this pipe has experienced the new failure within three months.



A, B, C, D, E, and F: denotes individuals in the study;
X: denotes an individual that experience the event.

Figure 3-1: Right censoring.
Adapted from: (Kleinbaum & Klein, 2005)

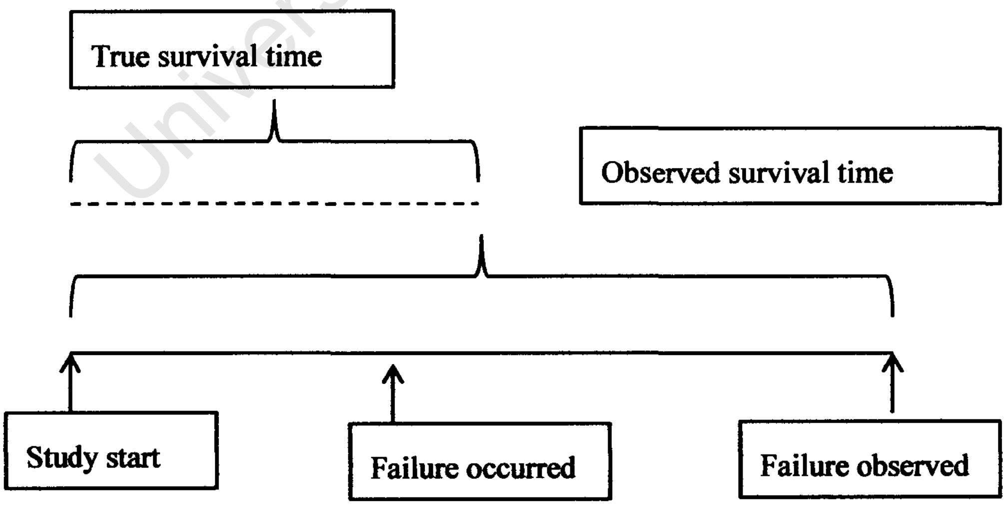


Figure 3-2: Left censoring.
Source: (Kleinbaum & Klein, 2005)

3.3 Descriptive Methods for Survival Data

The distribution of survival times is usually described or characterized by three functions: (1) the survivorship function, (2) the probability density function, and (3) the hazard function. These three functions are mathematically equivalent—if one of them is given, the other two can be derived. In practice, the three functions can be used to illustrate different aspects of the data. A basic problem in survival data analysis is to estimate from the sampled data one or more of these three functions and to draw inferences about the survival pattern in the population" (Lee & Wang, 2003).

In this section the different functions that describe the survival time distribution will be discussed as well as the different methods used to estimate them.

3.3.1 Survivor Function

The survivor function, also known as the survivorship and survival function, denotes the actual survival time of an individual, t, regarded as the value of a random variable T, which can take any non-negative value. The survivor function, denoted by $S(t)$, is defined as the probability that an individual survives longer than t, (Borovkova, 2002; Collett, 2003; Kleinbaum & Klein, 2005; Lee & Wang, 2003; Hosmer & Lemeshow, 1999).

$$\mathrm{S(t) = P(an\ individual\ survives\ longer\ than\ t) = P(T} > t). \qquad \text{Equation: 3.1}$$

In terms of the cumulative distribution function $F(t)$ of T, that is the probability that an individual fails before t, the survivor function is given by (Nathabandu & Rosso, 2008):

$$\begin{aligned} S(t) &= 1 - P(an\ individual\ fails\ before\ t), \\ &= 1 - P(T < t), = 1 - F(t). \end{aligned} \qquad \text{Equation: 3.2}$$

here $S(t)$ is a non-increasing function of time t with the following properties:

$$S(t) = \begin{cases} 1 & for\ t = 0 \\ 0 & for\ t = \infty' \end{cases} \qquad \text{Equation: 3.3}$$

The probability of surviving at least at the time zero is 1 and that of surviving an infinite length of time is zero. A typical plot for the survivor function, sometimes called the survival curve (Lee & Wang, 2003) , would be of the form illustrated in Figure 3-3.

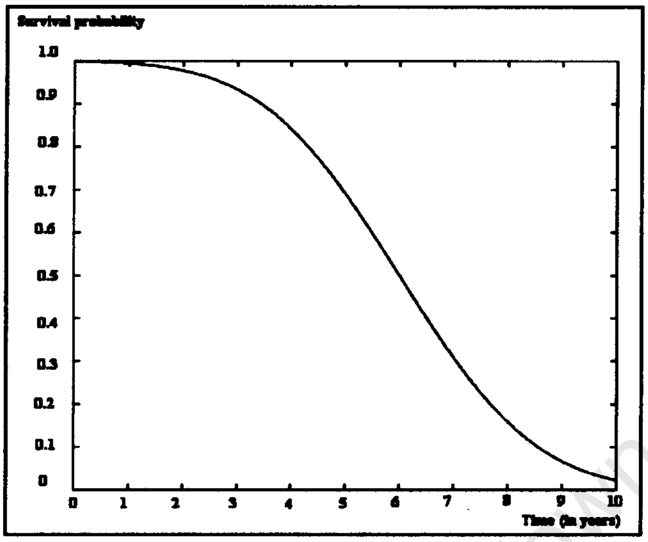


Figure 3-3: The survival function for a hypothetical population
Source: (Borovkova, 2002)Probability Density Function

The probability density function, also known as the unconditional failure rate, for a continuous random variable (e.g. survival time) is a nonnegative mathematical function that in its graphical representation usually takes the form of a continuous curve over a range of values that the random variable can possibly take. The probability density function, $f(t)$, of the random variable survival time is defined as the limit of the probability that an individual fails in a short interval t to $t + \Delta t$ per unit width Δt , or simply the probability of failure in a small interval per unit time, (Nathabandu & Rosso, 2008; Lee & Wang, 2003; Collett, 2003). It can be expressed as:

$$f(t) = \lim_{\Delta t \to 0} \frac{P\big(an\ individual\ fails\ in\ the\ interval\ (t, t + \Delta t)\big)}{\Delta t},$$
$$= \lim_{\Delta t \to 0} \frac{F(t+\Delta t)-F(t)}{\Delta t} = F'(t). \qquad \text{Equation: 3.4}$$

The graph of $f(t)$ is called the density curve. It has the following properties:

1. $f(t)$ is a non-negative function:

$$\begin{cases} f(t) \geq 0 & for\ all\ t \geq 0 \\ \quad\ = 0 & for\ t < 0 \end{cases}. \qquad \text{Equation: 3.5}$$

2. The area between the density curve and the t axis is equal to 1. (Lee & Wang, 2003)

Examples of the probability density function are given in Figure 3-4. The shaded areas between the density curve and the time axis represents the proportion of individuals failed between 1 and 2 units of time.

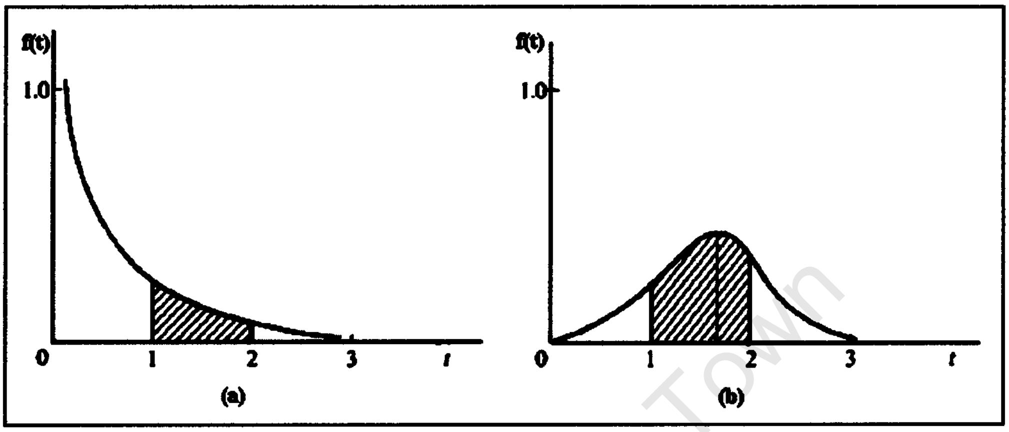


Figure 3-4: Examples of density curves.
Source: (Lee & Wang, 2003)Hazard Function

The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. In terms of the survival time T, it is defined as the limit of the probability that an individual fails in a very short interval, $t + \Delta t$, given that the individual has survived to time t. In contrast with the survivor function which focuses on not failing, the hazard function focuses on failing, i.e. the event occurs (Kleinbaum & Klein, 2005). The hazard function is also known as the instantaneous failure rate, the intensity rate, the force of mortality, the conditional failure/mortality rate, and age-specific failure rate (Lee & Wang, 2003; Collett, 2003). It can be expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\begin{pmatrix} \text{an individual fails in the time interval}(t, t+\Delta t), \\ \text{given the individual has survived to } t \end{pmatrix}}{\Delta t},$$

$$= lim_{\Delta t \to 0} \frac{P\{t \le T < t + \Delta t | T \ge t\}}{\Delta t}. \qquad \text{Equation: 3.6}$$

Some useful relationships between the survivor and hazard functions can be obtained from the above definition of the hazard function. Based on a standard result from probability theory, the probability of an event A, conditional on the occurrence of an event B, is given by (Collett, 2003; Nathabandu & Rosso, 2008):

$$P(A|B) = \frac{P(AB)}{P(B)}. \qquad \text{Equation: 3.7}$$

Where $P(AB)$ is the probability of the joint occurrence of A and B. By combining this result; the definitions of the cumulative distribution function and the survivor function; the conditional probability in the hazard function can be expressed as:

$$\frac{P(t \leq T \leq t+\Delta t)}{P(T \geq t)} = \frac{F(t+\Delta t)-F(t)}{S(t)}, \qquad \text{Equation: 3.8}$$

Now the hazard function can be written as:

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{F(t+\Delta t)-F(t)}{\Delta t} \right) \times \frac{1}{S(t)},$$
$$= \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}[\log S(t)]. \qquad \text{Equation: 3.9}$$

and so the probability of surviving to (t) becomes

$$S(t) = exp\left[-\int_0^t h(u)\, du\right] = exp[-H(t)], \qquad \text{Equation: 3.10}$$

The function $H(t)$ is known as the cumulative hazard function or the integrated hazard function. The cumulative hazard function can be obtained from the survivor function as:

$$H(t) = -\log S(t), \qquad \text{Equation: 3.11}$$

The hazard function may increase, decrease, remain constant, or indicate a more complicated process. Figure 3-5 shows a plot of several kinds of the hazard function.

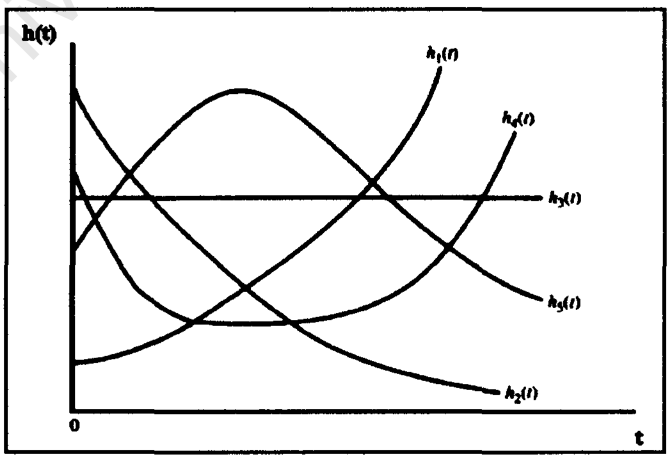


Figure 3-5: Examples of the hazard function.
Sources: (Lee & Wang, 2003)

In the analysis of survival data, the survivor function and the hazard function are estimated from the observed survival times. In section 3.4, methods that have been used to estimate the survival function will be presented.

3.4 Estimation of Survival Function

This section reviews the different methods used for the estimation of the three descriptive functions for a typical survival data set discussed in section0). There are generally three distinct methodologies for the estimation of survival functions; namely non-parametric methods, parametric methods and semi parametric methods. The three methods are described as (Shoukri & Pause, 1999; Hosmer & Lemeshow, 1999 Kalbfleisch & Prentice, (2002); Borovkova, (2002); Collett, (2003); Nelson, (2003); Jenkins, (2005)).:

- **Non-parametric or distribution free methods**: This approach does not require specific assumptions about the underlying distribution of the survival times. Non-parametric methods that are mostly used to obtain such estimates are:
  - The Life-table estimator

    The life-table estimator, also known as the Actuarial estimator of survivor function, has been used for more than 100 years to describe human mortality experience and is among the earliest examples of the application of statistical methods. It is obtained by first dividing the period of observation into a series of time intervals which depend on the number of individuals in the study.

    Define intervals of time I_j where $j = 1, \dots, J$: $I_j : [t_j, t_{j+1})$, where

    d_j: the number of failures observed in interval I_j

    m_j: the number of censored spell endings observed in interval I_j

    N_j: the number at risk of failure at start of interval

    n_j: the adjusted number at risk of failure used for midpoint of interval

    $$n_j = N_j - \frac{d_j}{2}. \qquad \text{Equation: 3.12}$$

    and hence

    $$S(j) = \prod_{k=1}^{j}\left(1 - \frac{d_k}{n_k}\right). \qquad \text{Equation: 3.13}$$

an estimation of the density function can then be derived as (recall $S(t) = 1 - F(t)$)

$$f(j) = \frac{F(j+1) - F(j)}{t_{j+1} - t_j} = \frac{S(j) - S(j+1)}{t_{j+1} - t_j}. \qquad \text{Equation: 3.14}$$

and an estimate of the hazard rate is given by,

$$H(j) = \frac{[f(j)]}{\tilde{S}(j)}, \qquad \text{Equation: 3.15}$$

where

$$\tilde{S}(j) = \frac{S(k) + S(k+1)}{2} \qquad \text{Equation: 3.16}$$

and is taken as applying to the time corresponding to the midpoint of the interval

- The Kaplan-Meier Estimator

In contrast with the life-table estimator which assumes grouped survival times, this estimator assumes continuous time measure of spells. Consider the following quantities in a data set:

d_j: the number of persons observed to fail (make transition out of the state) at t_j

m_j: the number of persons whose observed duration is censored in the interval $[t_j, t_{j+1})$, i.e. still in state at time t but not in state by $t + 1$

n_j: the number of persons at risk of making a transition (ending their spell) immediately prior to t_j, which is made up of those who have a censored or completed spell of length t_j or longer:

$$n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + \cdots + (m_k + d_k). \qquad \text{Equation: 3.17}$$

The Kaplan-Meier estimate of the survivor function at survival time t_i is given by the product of one minus the number of persons who exits survival time t_i divided by the number of persons at risk of exit,

$$S(t_j) = \prod_{j|t_j<t}\left(1 - \frac{d_j}{n_j}\right),$$ Equation: 3.18

Thereafter estimates of the density and hazard functions can be obtained.

- The Nelson-Aalen or Altshuler's Estimator

  Another alternative estimate of the survivor function is the Nelson-Aalen estimate. In contrast with the Kaplan-Meier estimate, his estimate is obtained from an estimate of the cumulative hazard function, given by

  $$H(t_j) = \sum_{j|t_j<t}\left(\frac{d_j}{n_j}\right).$$ Equation: 3.19

  This provides an estimate of the survivor function which is $\exp\left(-\mathrm{H}(\mathrm{t_j})\right)$.

Another nonparametric estimator for survival data is the **Nelson's Mean Cumulative Function (MCF) Estimator** by (Nelson, 2003). This estimator is different from the three former estimators in that it provides estimates for the mean cumulative function $M(t)$ of the number of events rather than the survivor of the hazard functions. In most applications, $M(t)$ is an increasing function of age (t) and it yields most of the information sought from recurrence data such as recurrence or failure rate, prediction, comparisons, etc.

A detailed description of the above methods for non-parametric estimation for survival function is beyond the scope of this thesis; however a brief description of the last method developed by Nelson, (2003) will be presented in section 4.2. The method will be used to obtain estimates for the failure rate and the number of failures for the case study water pipelines.

- **Parametric methods**: all parametric methods involve specification of a distributional form for the probability density function $f(t)$. This in turn specifies the survival function and the hazard function using the relationships defined previously. There are a number of parametric models that have been used in survival analysis, some of them are (Shoukri & Pause, 1999; Collett, 2003; Kalbfleisch & Prentice, 2002):

- The Exponential Model

The simplest model for the hazard function is to assume that it is constant over time. Under this model, the hazard function can be expressed as:

$$h(t) = \lambda, \quad for\ 0 \leq t < \infty . \qquad \text{Equation: 3.20}$$

The parameter λ is a positive constant which would be estimated by fitting the model to the observed data. From the above equation, the corresponding survivor function is

$$S(t) = exp\left[-\int_0^t \lambda\, du\right] = e^{-\lambda t} . \qquad \text{Equation: 3.21}$$

And so the implied probability density function of the survival time is

$$f(t) = \lambda e^{-\lambda t}, \quad for\ 0 \leq t < \infty . \qquad \text{Equation: 3.22}$$

Note that the hazard function is independent of time, implying the instantaneous conditional failure rate does not change within a lifetime. This is also referred to as the memory-less property of the Exponential Distribution, since the age of an individual does not affect the probability of future survival. When $\lambda = 1$, the distribution is referred to as the unit exponential. Figure 3-6 illustrates the functions $h(t)$, $S(t)$, and $f(t)$ of the Exponential Distribution.

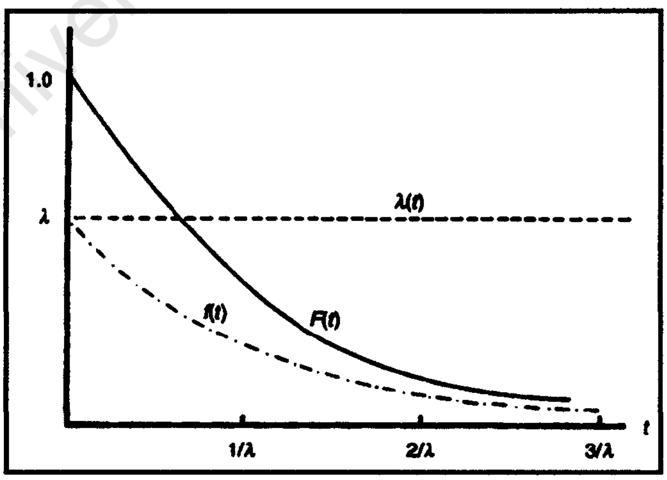


Figure 3-6: Hazard function, density function, and survivor function for the single parameter Exponential Distribution, note that λ may exceed 1
Source: (Kalbfleisch & Prentice, 2002)

- The Weibull Model

An important generalization of the Exponential Model allows for a power dependence of the hazard on time. This yields the two-parameter Weibull distribution with a hazard function with the form:

$$h(t) = \lambda(t) = \lambda\gamma(\lambda t)^{\gamma-1} \quad for\ \lambda, \gamma \geq 0; t \geq 0\,. \qquad \text{Equation: 3.23}$$

This hazard is monotone decreasing for $\gamma < 1$, increasing for $\gamma > 1$, and reduces to the constant hazard (i.e. the Exponential Model), if $\gamma = 1$. The probability density function is

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\ exp[-(\lambda t)^{\gamma}]\,. \qquad \text{Equation: 3.24}$$

and the survivor function is

$$S(t) = exp[-(\lambda t)^{\gamma}]\,. \qquad \text{Equation: 3.25}$$

The shape of the hazard function depends critically on the value of γ, and so γ is known as the shape parameter, while the parameter λ is a scale parameter. Figure 3-7 illustrates the hazard function h(t) with different values of γ for the two-parameter Wiebull Distribution.

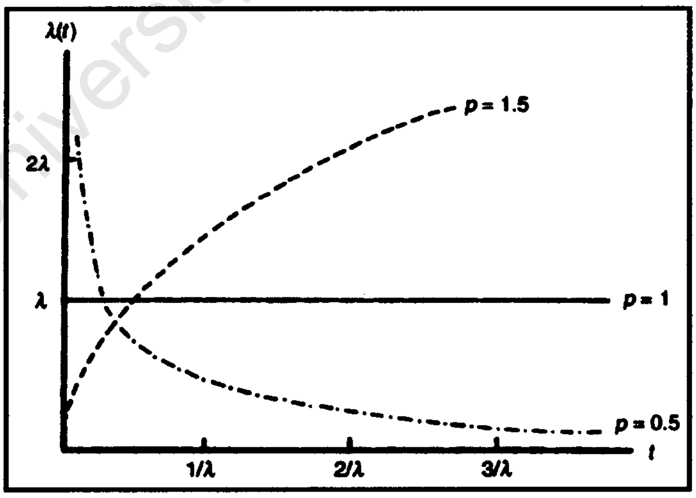


Figure 3-7: Hazard functions for the two-parameter Wiebull Distribution with shape parameter $\gamma = P$)

Source: (Kalbfleisch & Prentice, 2002)

Other examples of parametric models are, (Lee & Wang, 2003; Hosmer & Lemeshow, 1999; Kalbfleisch & Prentice, 2002):

- The Log-Normal Distribution
- Gamma and Generalized Gamma Distributions
- Log-Logistic Distribution
- Generalized F Distribution

- **Semi-parametric methods**: the most representative model for these methods is the "Relative Risk" or "Cox Regression model" (Kalbfleisch & Prentice, 2002). This model constitutes of both parametric and non-parametric parts and therefore it is referred to as semi-parametric model. The non-parametric aspect is represented by an unspecified baseline hazard function which is a function of time. The parametric aspect incorporates parametric modeling of the hazard rate and a specified set of explanatory variables that are fixed or time independent. However, it is also possible to consider explanatory factors which are time dependent but with different characteristics. In this case, the model is referred to as the Extended Cox Model (Collett, 2003; Hosmer & Lemeshow, 1999; Kalbfleisch & Prentice, 2002; Kleinbaum & Klein, 2005; Machin, Cheung, & Parmar, 2006). The Cox's model will be discussed in details in chapter 4 as a method of modeling pipe failure in this research.

- **Accelerated Lifetime models**

  In studies of mortality, it usually makes sense to think in terms of the hazard and the hazard ratio when making comparisons between groups. In some other situations, it is easier to think in terms of the relative length of time to the event. In other words, the time-to-event is accelerated. The corresponding models are termed Accelerated Failure Time (AFT) models. (Machin, Cheung, & Parmar, 2006)

The motivation behind the use of survival analysis as a method of pipe failure modeling in this dissertation is discussed in the next chapter. Survival analysis has been found to provide direct estimates of the failure probability rather than the annual expected number of failures. An important property of the hazard model reported in the literature was the flexibility of the model in terms of the estimation of the baseline hazard function and the coefficients of the explanatory variables. This property allows for the independent capture of the ageing process of pipes and the effect of the other included explanatory factors. However the main problem that remains is the availability of the suitable and sufficient data for this type of analysis.

In chapter 4, the two survival methods that have been used in the analysis of the failure of the 100 mm FC pipes in Cape Town will be discussed in details.

4 Applied Survival Methods to the Case Study

4.1 Introduction

In section3.4, some of the methods used to estimate survival functions have been presented. Several methods have been described for parametric and non-parametric estimations. The focus in this chapter will be in describing the two statistical survival methods which will be used in this thesis in the analysis of the case study pipeline failures namely, the Nelson Mean Cumulative Function Estimator and the Semi-Parametric Cox's Proportional Hazard Model.

4.2 The Nelson Mean Cumulative Function Estimator

The population Mean Cumulative Function (MCF), firstly developed by Nelson (1988), for the number or cost of repairs per system as a function of population age, described in section 3.4, is a very informative estimate. In reliability application, a plot of the MCF versus age can be used to, (Nelson, 1995)

- Evaluate how a cost or repair rate increases or decreases with system age;
- Compare different samples from different designs, production periods, maintenance; policies, environments, operating conditions, etc;
- Predict future numbers and costs of repairs; and
- Reveal unexpected infromation and insight.

4.2.1 Definition of the MCF

"*At any age t, the corresponding distribution of the number or cost of events has a mean* $M(t)$. *This mean as a function of t is called the mean cumulative function (MCF), see Figure (4-1). This function can be regarded as the "mean curve," as it is the point wise average of all population curves passing through the vertical line at each age. For many applications, the mean curve is regarded as continuous. In most applications,* $M(t)$ *is an increasing function of age t. For costs, it is called the MCF for costs (of events). For the number of events, it is called the MCF for the number of events.*" (Nelson, 2003)

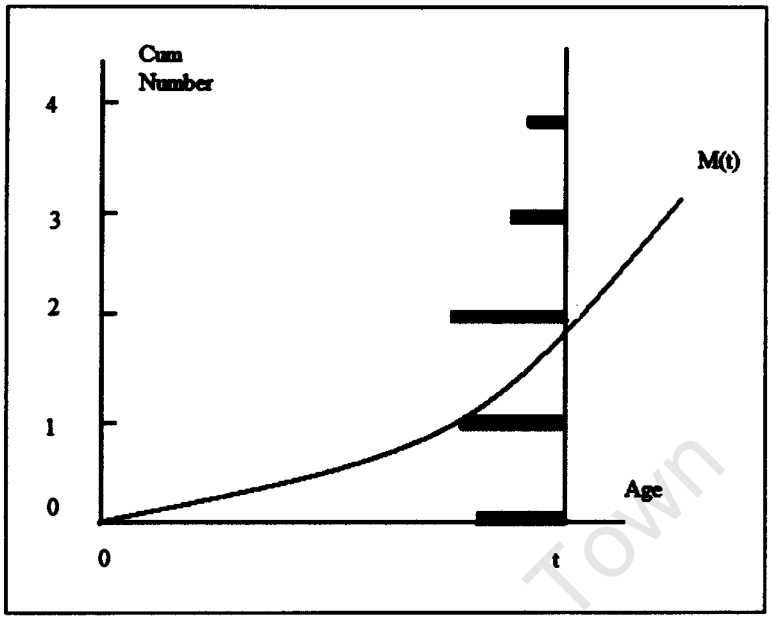


Figure 4-1: Discrete distribution of cumulative number of recurrences at age t
Source: (Nelson, 2003)

One of the most important information sought from this estimate, and one of the main goals in this work, is the Failure Rate $m\ (t)$ which is called the instantaneous population recurrence rate or intensity function. The failure rate is expressed in the number or cost of events per month (hour, year, mile, etc.) per population unit and, for count data, is given by the derivative, (Nelson, 2003)

$$m\ (t) = \frac{dM(t)}{dt} \qquad \text{Equation: 4.1}$$

Nelson, (2003), presented a variety of methods by which this function is estimated from both exact age data with different censoring statuses and continuous history function data. The method of concern in this thesis is how to calculate a nonparametric estimate of the population MCF from exact age data with right censoring which will be discussed in details in section 5.3.2.5. Other possible information sought from the MCF and the methods used for their estimation can be found in (Nelson, 2003).

4.2.2 Advantages and Disadvantages of the MCF Plots

Nelson, (2003), reported some advantages and disadvantages of the data plots of his MCF estimates and these are

- Advantages of data plots
  - Plots are simple, quick to make, and easy to interpret. Common statistical packages make such plots, and many plots are easy to make with a spreadsheet.
  - They provide an estimate of the population MCF and other quantities of interest.
  - They allow one to assess how well a proposed parametric model fits the data.
  - They help convince others of conclusions based on the plots or on analytic methods.
  - They reveal unsought insights into the data.

- Disadvantages. Plots have the following disadvantages:
  - The accuracy of estimates from plots is unknown, although some experienced analysts can subjectively judge accuracy. It is best to also use analytic methods to calculate and plot confidence limits, which are objective. Inexperienced analysts tend to think estimates are more accurate than they actually are.
  - Graphical comparisons of data sets may be inconclusive unless sets differ greatly. Such comparisons are more accurate if aided by confidence limits or hypothesis tests.
  - Plots cannot be used to determine appropriate sample sizes.

4.3 The Cox's Proportional Hazard Model (PHM)

The Cox's proportional Hazard model (PHM) developed by (Cox, 1972) is a survival regression model and therefore it is sometimes called the Cox's regression model. In terms of some characteristics of the model, it is sometimes referred to as the semi-parametric Cox's model, the relative hazard model or simply the proportional Hazard model, (Kalbfleisch & Prentice, 2002; Collett, 2003; Kleinbaum & Klein, 2005; Machin et al., 2006). The Cox's model is well recognized in various research fields as one of the most popular statistical models to describe lifetime distribution for individuals. The model has been extensively used in medical, economical and reliability engineering areas of research.

The description of the PHM in this section will start with a general presentation of the mathematical formula of the model, its different statistical inferences, the common reason of the popularity of the model, and the procedures of model building. A description of the application of the model in water pipeline failures will be discussed in a separate section in chapter 5.

4.3.1 Characteristics of the Cox' Proportional Hazard Model

In the Cox's PHM, the hazard rate for an individual is expressed as a function of an unspecified baseline hazard function and a vector of covariates related to that individual. The baseline hazard function denotes the hazard function for the individual as a function of time (i.e. the ageing process). The general form of the PHM is:

$$h(t,X) = h_0(t)\exp\left(\sum_{i=1}^{p} \beta_i x_i\right), \qquad \text{Equation: 4.2}$$

where:

$h(t,X)$ The hazard function for the i_{th} individual,

X Set of explanatory variables, $X_1, X_2, X_3, \ldots, X_p$

x_i Values of i_{th} individual's set of explanatory variables, $x_1, x_2, x_3, \ldots, x_p$

h_0 The baseline hazard function

p Number of individuals

β_i Regression parameters associated with x_i

An important feature of this model is that the baseline hazard function, the time dependent, component does not involve covariates, and the exponential term is time independent. Therefore, in cases where no explanatory variables are considered, the baseline hazard function can be interpreted as the hazard function of a component with a set of covariates equal to 0. In other words, the Cox's model reduces to the baseline hazard function (Kleinbaum & Klein, 2005; Machin et al., 2006).

However, it is also possible to accommodate time dependent covariates in Cox's model under different specifications and assumptions through a modified model called the Extended Cox's Model. The other important feature of the model is that the baseline hazard function is unspecified. That is, there is no particular mathematical form is specified for it and therefore the model is semi-parametric (Kleinbaum & Klein, 2005; Machin et al., 2006; Collett, 2003; Kalbfleisch & Prentice, 2002). The PHM in Equation 4.2 can be expressed as

In order to examine the dependency of the hazard function in the PHM on time, the graphical plot of the complementary log transformation, or $log\,(-log)$ survival function $s(t)$, against log of the failure time (t), may be used. Linear relationship in the plot indicates that the hazard function is constant, while deviation from linearity indicates that the hazard function is changing over time. The survival function in this test may be estimated using the non-parametric methods described in section 3.4, such as the Kaplan-Meier estimate (Collett, 2003; Machin et al., 2006).

4.3.6.2 Tests for Proportional Hazards

It has been indicated earlier that the hazard functions under the PHM for two individuals with different values of explanatory variables are assumed to be proportional and independent of time. This property of the PHM is expressed in terms of the hazard ratio $\widehat{HR}$ in (Kleinbaum & Klein, 2005) as follows

$$\widehat{HR} = \frac{\hat{h}(t,X^*)}{\hat{h}(t,X)} = \frac{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}} = exp\left[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right] \qquad \text{Equation: 4.15}$$

Where X^* and X denote the explanatory variables for the two individuals. Therefore, once the regression parameters are obtained from the fitted model, the hazard ratio can easily be computed by substituting the values of parameters and explanatory variables in Equation 4.15.

This assumption may be tested by three major statistical methods namely the graphical methods, the goodness of fit approach and the time dependent covariates approach. One of the graphical methods is an extension of the log transformation test for constant hazards described above. The other method is a plot of observed vs. expected failures, (Kleinbaum & Klein, 2005).

The goodness of fit test distinguishes from the graphical methods in that it provides a test statistic and p-value for assessing the proportional hazards assumption for a given predictor of interest. It allows, therefore, for a more objective decision using a statistical test than when the graphical methods are used, (Kleinbaum & Klein, 2005). The goodness of fit will be used in this thesis for checking the proportionality assumption for different explanatory variables. Further description of the test will be provided in chapter6.

The third method for examining the proportional hazards assumption in the PHM is based on the fact that if any of the explanatory variables are varying with time the proportional hazard assumption is violated, (Lee & Wang, 2003). The test is incurred by adding a time-dependent variable to the model or by including an interaction term involving the time-independent variable being assessed and some function of time. The proportional hazards assumption is examined by testing for the significance of the interaction term where the predictors are to be assessed one-at-a-time. The test can be performed by using either a Wald statistic or a likelihood statistic. In either case the statistic has a chi-square distribution with one degree of freedom under the null hypothesis. The proportional hazards assumption can also be examined for several predictors simultaneously and for a given predictor adjusted for other predictors using the time-dependent variable method. Details about this method can be found in Kleinbaum & Klein, 2005, Collett, 2003 and Machin et al., (2006).

"*However, the violation of this assumption is not necessarily unacceptable. To assume proportional hazards is essentially to estimate the effect of an independent variable averaged over time and ignore the possibility that the effect may vary over time. A simple model based on the assumption of proportional hazards is sometimes preferable to a more 'correct' but complex model if the added complexity does not identify features of scientific or clinical significance*," (Machin et al., 2006).

4.3.6.3 Confidence Intervals and Hypothesis Tests for Regression Parameters

When a statistical package is used to fit a PHM, the provided parameter estimates are usually accompanied by their standard errors which can be used to obtain approximate confidence intervals for the unknown regression parameters. Particularly, a $100(1-\alpha)\%$ confidence interval for a parameter β is the interval with limits $\hat{\beta} \pm z_{\alpha/2} s.e.(\hat{\beta})$, where $\hat{\beta}$ is the estimate of β and $z_{\alpha/2}$ is the upper $\alpha/2$ -point of the standard normal distribution. If a $100(1-\alpha)\%$ confidence interval for β does not include zero, this is evidence that the value of β is non-zero. Furthermore, the null hypothesis that $\beta = 0$; can be tested by calculating the value of the statistic $\hat{\beta}/s.e.(\hat{\beta})$. The observed value of this statistic is then compared to percentage points of the standard normal distribution in order to obtain the corresponding P-value. On the other hand the square of this statistic can be compared with percentage points of a chi-squared distribution on one degree of freedom. In most statistical packages the P-values for this test are often given beside parameter estimates and their standard errors in computer output, (Collett, 2003).

4.3.7 Stratified Cox's Model (SCM)

The Stratified Cox's Model, SCM, is a modification of the PHM that allows for control by stratification of a variable in the model that does not satisfy the proportional hazards assumption. Variables that are assumed to satisfy the PH assumptions are included in the model; whereas the stratified variable is not included. A generalization of the SCM allows for stratification for several variables over several strata, (Kleinbaum & Klein, 2005).

The SCM is best used in cases where the stratifying variables are known to be potential confounders but their effects on the outcome are not of direct science interest, or of secondary importance to those of other variables, (Machin et al., 2006; Hosmer & Lemeshow, 1999).

The following is a description of the general form of the SCM which has been provided by (Kleinbaum & Klein, 2005). Assume a set of explanatory variables with k variables not satisfying the PH assumption and p variables that are satisfying the PH assumption. Denote the variables not satisfying the PH assumption as $Z_1, Z_2, Z_3, \dots, Z_k$; and the variables satisfying the PH assumption as $X_1, X_2, X_3, \dots, X_p$.

To perform the SCM, define a single variable Z^* from the Z group of variables to be used for stratification. This is done by forming categories of each Z_i and then form combinations of categories, these combinations will form the strata. These strata are the categories of the new variable Z^*.

The general form of the SCM can then be expressed as

$$h_g(t,X) = h_{0g}(t)exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p] \qquad \text{Equation: 4.16}$$

where

$g = 1, 2, \dots, k^*$, Strata defined from Z^*

k^* is the total number of combination (or strata) formed after categorizing each of the Z variables.

Note that although the coefficients $\beta_1, \beta_2, \dots, \beta_p$ in the SCM are the same for each stratum, the baseline hazard function is allowed to be different for each stratum. However, because of the regression coefficients are the same for each stratum estimates of the hazard ratio will be the same for each stratum.

Estimates of the $\beta's$ can be obtained by maximizing the partial likelihood function which is the product of the partial likelihood functions for each stratum. The partial likelihood function for the SCM has been expressed in (Kalbfleisch & Prentice, 2002) as

$$L(\beta) = \prod_{j=1}^{q} L_j(\beta) \qquad \text{Equation: 4.17}$$

where $L_j(\beta)$ is the partial likelihood function for β arising from the j_{th} stratum alone. The same maximization procedure of the PHM is applied for maximizing the partial likelihood function for the SCM. Similarly, estimates of the baseline hazard function can be obtained using the same methods as those for the PHM. Note that the partial likelihood function for each stratum is derived from its corresponding hazard function (Kleinbaum & Klein, 2005).

4.3.8 Experiences with Survival Analysis in Pipe Failure Modeling

This section discusses the experiences of the utilization of survival analysis in modeling failures of pipes in water distribution networks. This includes a review of the proposed model, the methodologies used in models development and a summary of the outcomes from the applied models and methodologies.

The application of survival analysis in modeling water pipes failure dates back to the 1980s. The first attempt was made by (Jeffrey, 1985) where a PHM was applied to the water distribution network break data in New Haven city, United States. The aim of the study was to support decision making for water distribution management through the provision of necessary information on the projected status of the network at different points of time in the future, the cost and the risks associated with alternative maintenance programs, as well as information on the probability of failure of each pipe. A PHM was developed for predicting failure for each individual pipe in the system and contributory factors leading to failure were identified. It was concluded that the PHM can be successfully used to support decision making of the appropriate maintenance management strategies of water distribution systems.

This has been followed by a study conducted by (Andreou S. A., 1986) to study failure patterns and probabilities of failure in two case studies of water distribution systems break data in the United States. The focus of this study was to capture the high variability in break rates that exists among individual pipes within and between different systems; and to identify the related predicting variables that lead to pipe failure.

The first case study was the same case study analysed by (Jeffrey, 1985). The applied methodologies identified three stages of deterioration of water pipes in the second system during its useful life. A PHM was proposed to describe the early stage of pipes deterioration. This was motivated by the fact this stage incurred only few frequent failure, thus, predicting the probability of failure was found to be more reasonable than predicting the failure rate. Estimates for breakage rates were obtained for the second stage of deterioration in which a multiple and frequent breaks were experienced. This stage has been described by an exponential type regression model and future breakages were represented as Poisson arrival. A PHM was also applied to determine the probability of entering the fast stage of deterioration. Again the integration of the proposed model in making maintenance decisions was examined.

A comparison between the two system performances was held and results showed that the first system is generally in a better condition than the second system. Theory behind the applied models and their application on the case study were published in (Andreou, 1987) and (Andreou, 1987) respectively.

The methodologies used in developing the PHMs in the two mentioned studies are similar to the methodologies described in the previous sections and a detailed description of the models building procedures is presented in Table 4-1.

Another experience with survival analysis modelling of pipe failure is the models developed by (Gustafson & Clancy, 1999). In this study the deterioration of cast iron water mains was modelled as semi-Markov process. Pipes in the case study network were categorized according to the wall thickness and year of installation of pipes. Estimates of the probability distribution of failure times were obtained using several methods of survival analysis. The distribution of time to first break was assumed to follow a Generalized Gamma distribution, whereas an Exponential model was proposed to describe the distribution of time to second and subsequent breaks. Finally the effect of contributory factors was examined using an Accelerated Lifetime model. The three models were checked for adequacy and confidence intervals for the three fitted distributions.

The authors used the probability distributions with parameters to forecast future breaks in cast iron water mains in the case study network, and the models were calibrated against the actual performance of the system by the mean of a semi-Markov process. The methodologies and variables used in models development are presented in Table 4-1.

Two statistical models have been developed by (Røstum, 2000) for a case study of water main break database. The purpose of the study was to examine these models for predicting failure for each pipe in the case study network and to determine whether the existing data is sufficient input for these models or not. The first model is the parametric Weibull PHM. The second approach includes the use of a counting process (Non Homogeneous Poisson Process, NHPP). Pipes in the network were categorized into several groups based on different failure characteristics.

The following variables were found to be significant: pipe length, pipe diameter, soil condition, pipe age at the time of failure records starts and the number of previous failures. These variables varied in their significance from group to group; and from strata to strata in the case of the Weibull PHM. The models were calibrated using the existing failure data of a nine year period, verified with data of the following two years and finally used in the prediction of breaks at both pipe and network level.

The two models were evaluated and found to be capable to model failures of pipes in the network, however, an over estimation tendency was reported to the Weibull PHM. Even though the model has had the advantage of being able to model pipes where the covariates had a positive effect and where several breaks had already occurred. In general, the NHPP model was recommended based on the results from case study in a manner that it provides better predictions of pipes failure than the Weibull PHM. Predictions of the expected number of failures has been obtained and proposed to enhance decision making process in the water industry. A directed collection of detailed water mains break data was recommended for more accuracy in developing predicting models for pipe failure.

A PHM was applied by Park, (2004) to a case study water main break database to identify the hazard characteristics in relation to the breaks causing factors. A soil survey was conducted in order to obtain information on the characteristics of the soil and level of land development in the case study area, which have been used as covariates in the PHM. Based on the results from the applied PHMs, the critical point was found to be the second break in cast iron pipes; and the third break for the ductile iron pipes.

Pipes in the data set were categorized into 8 groups according to material, size, and the number of previous breaks before the last break for various categories of pipes. Results from the models were interpreted so as an insight on the future system performance can be obtained in order to plan for the appropriate management strategies for the network. To enhance the capability and the usefulness of the PHM in characterizing failures in water distribution networks pipes, the paper recommended the acquisition of an informed break data base. Results from the model were considered to serve as guidelines on which category of pipes needs more preventative rehabilitation or replacements.

More recently, a study by (Park S. et al., 2010 used the PHM to model the time between consecutive pipe breaks using a case study water main break data. During the process of building the PHMs, the assumption of the proportional hazards was examined. It was found that some of the categories in the data set have one or two covariates that have time-dependent effects on the hazard rates. The models were then extended to capture the effect of these time-dependent variables.

By analyzing the baseline hazard rates, the hazards of the third through the seventh break were found to follow a form similar to the bath-tub curve. The collection of more datialed data was recommended by the authers in order to direct future utilization of the PHM in making management decisions for water networks. More specifically, the collection of data such as pressure, traffic volume and soil characteristics was requierd for the construction of the PHM in order to be able to efficiently allocate funds so that the hazards of pipe failures are reduced. It was also referred to the usefulness of the estimated survival functions of a pipe in the provision of insight about the general conditions of a pipe of interest if subsequent breaks are assumed to occur.

A summary table of other survival analysis models that have been used in modelling water pipes failure in the literature are presented in Table 4-1.

Table 4-1: Literature review of survival pipe failure models

Reference	Model equation	Notation	variables	Methodology
(Jeffrey, 1985)	• Proportional Hazard: • General form of the Proportional Hazard Model: $h(T,X) = h_0(t)e^{\sum_{i=1}^{j} B_i X_i}$ • New Haven system (baseline hazard function): • $h_0(t) = 2\times10^{-4} - 10^{-6}t + 2\times10^{-7}t^2$	T = time to next break $h(t,X)$ = hazard function $h_0(t)$ = baseline hazard function X = vector of explanatory variables B = vector of coefficients to be estimated t = survival time since last break	• Data set variables: ▪ Diameter, ▪ Length, ▪ Pressure, ▪ Pipe type, ▪ Soil corrosivity, ▪ Soil stability, ▪ Land development, ▪ Swamp, ▪ Installation date, ▪ Number of previous breaks, ▪ Time to first repair • Model variables: ▪ Natural logarithm of length, ▪ Pressure in 10 psi, ▪ Installation period, ▪ Age of pipe at the last break, ▪ Number of previous breaks	• Preliminary Statistics to identify the range and variability of variables in the data set; • Bivariate Analysis to identify any possible correlation between pipe failure and other variables in the data set (three types of correlation); • Grouping of data in several strata • Treatment of left censoring problem; • Survival Analysis to reveal survival patterns for various categories of pipes; • Identification of the model variables; • Estimation of regression coefficients using Maximum Likelihood Method; • Estimation of the baseline hazard function; • Final configuration of the PHM • Software package used is BMDP

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Andreou, 1986)	Proportional Hazard: General form: Same as in (Jeffrey, 1985)	Same as in (Jeffrey, 1985)		
(Andreou, 1986)	Proportional Hazard: System (A), New Haven (baseline hazard function): $h_0(t) = 2 \times 10^{-4} - 10^{-6}t + 2 \times 10^{-7}t^2$	$h_0(t)$ = baseline hazard function t = Survival time since installation, if breaks have not occurred, or = time since last break if breaks have occurred	• Data set variables: ▪ Same as in (Jeffrey, 1985) • Model variables: ▪ Natural logarithm of length; ▪ Pressure in psi; ▪ Percentage of low land development ▪ covering the pipe; ▪ Installation period; ▪ Break rate at second break; ▪ No. of previous breaks (only if 2 breaks are reported)	Same as in (Jeffrey, 1985) Software package used is BMDP

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Andreou, 1986)	Proportional Hazard: System (B), Cincinnati: Baseline hazard function for time to second break conditional on one previous break: $h_0(t) = 0.1297 - 0.0086t + 2.10^{-4}t^2$	t = survival time since last break	• Natural logarithm of length; • Absolute internal pipe pressure; • Percentage of pipe in highly corrosive soil; • Pipe diameter; • Land use variables: • Percentage of residential land development covering the pipe; • Percentage of industrial land; • Percentage of commercial land; • Percentage of transportation land; • Population density; • Period of installation; • Several interaction variables; • Break rate at the last break; • Effect of lining and cleaning (only for time from first to second break and for break rate after the third and sixth models)	• Same as in New Haven system except for stratification of failure time as a result of non-satisfactory of the proportional hazard assumption which resulted in different PHM for each stratum; • Software package used is BMDP

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Andreou, 1986)	• Proportional Hazard: • System (B), Cincinnati: • Baseline hazard function for time to fast break stage conditional on one previous break: obtained by calculating the average numerical values in 5 year time interval • Exponential models for break rates after the third and sixth failure. General form: • $R = exp(Bz) + e$	R = the estimated yearly break rate B = the set of estimated coefficients z = the covariate vector e = the error term of the model	• Data set variables: ▪ Same as in system (B) above • Model variables: ▪ Land use variables: ▪ Percentage of residential land development covering the pipe; ▪ Percentage of industrial land; ▪ Percentage of commercial land; ▪ Percentage of transportation land; ▪ Population density; ▪ Period of installation; ▪ Several interaction variables; ▪ Break rate at the last break; ▪ Effect of lining and cleaning (only for time from first to second break and for break rate after the third and sixth models)	• Same as in New Haven system except for stratification of failure time as a result of non-satisfactory of the proportional hazard assumption which resulted in different PHM for each stratum; • Software package used is BMDP

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Gustafson & Clancy, 1999)	• Survival models: • Generalized Gamma Distribution for modeling the time to first break. The distribution of the Generalized Gamma Distribution used is given by: • $\frac{\beta}{\Gamma(k)} \frac{t^{\beta k-1}}{\alpha^{\beta k}} exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]$ •	• t = time • $k = \frac{1}{\lambda^2}$ • $\beta = \frac{1}{\sigma\sqrt{k}}$ • $\alpha = exp(\mu - \sigma\sqrt{k}\, logk)$ • And where: • $\mu =$ intercept • $\sigma =$ scale • $\lambda =$ shape	• Data set variables: ▪ ID number for each pipe ▪ Location by street, ▪ intersecting street and/or block numbers ▪ Pipe material ▪ Joint type ▪ Nominal diameter • Model variables: ▪ The exposed area as a surrogate for length	• Categorization of pipes according to year of installation, wall thickness, and tensile strength; • Modeling break history as a Semi-Markov process (i.e. the probability distribution for a failure time is independent of its previous holding time (age), but on its previous numbers of failure); • For each pipe category: • Time to first break: • A plot of number of pipe experiencing their first break vs. years since installation to identify observations and the sum of the probability distribution of the predicted failure times for the censors by the year of installation; • A plot of the Generalized Gamma Distribution to fit the observations and censors using the *LIFEBERG*[6] procedure for the statistical package SAS/STAT6; • The programme output the intercept, scale, and shape parameters of the model and the predicted Mean Time to First Break. Confidence intervals for the model are evaluated to determine the usefulness of the proposed distribution.

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Gustafson & Clancy, 1999)	• Survival models: • Exponential Model for time from first to second breaks and so on. The distribution of the exponential model used is given by: $\frac{1}{\theta} exp\left(-\frac{t}{\theta}\right) \quad t > 0$	t = time $\theta = exp(\mu)$ And, where: μ = intercept	• Data set variables: ▪ ID number for each pipe ▪ Location by street, ▪ intersecting street and/or block numbers ▪ Pipe material ▪ Joint type ▪ Nominal diameter ▪ Length ▪ Year of installation ▪ Break occurrence year ▪ Replacement year , if the pipe has been replaced • Model variables: ▪ The exposed area as a surrogate for length	Time to second break: Same procedures for the time to first break are used for the Exponential Distribution; The Lagrange Multiplier Chi Square indicator is used by the *LIFEBERG*[6] procedure of SAS/STAT6 to indentify the acceptance of the Exponential Distribution; Confidence intervals were tested, and the Mean Time to Second and subsequent failures are predicted;

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Gustafson & Clancy, 1999)	• Survival models: • Accelerated lifetime model: for the multiplicative effect of contributory variables in the time to break. For this case the time t_0 input to the Generalized Gamma distribution and the Exponential distribution is determined as follows: • $t_0 = t\,exp(-x'\beta)$ •	t = time x' = row vector of the values of the contributory variables β = column vector of parameters reported by $LIFEBERG^6$ for contributory variables	• Data set variables: ▪ ID number for each pipe ▪ Location by street, ▪ intersecting street and/or block numbers ▪ Pipe material ▪ Joint type ▪ Nominal diameter ▪ Length ▪ Year of installation ▪ Break occurrence year ▪ Replacement year , if the pipe has been replaced • Model variables: ▪ The exposed area as a surrogate for length	Effect of contributory factors: A variable named the "exposed area" was included in the analysis to examine the effect of length on failure times. The $LIFEBERG^6$ output the results for the significance of the variable "exposed area."

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Park S. , 2004)	Proportional Hazard: Same as in (Jeffrey, 1985)	Same as in (Jeffrey, 1985)		• Compilation of the Complete Break Database; • Soil survey to assign covariate values; • Assignment of soil association for each pipe ID in the Complete Break Database; • Development of a MATLAB programme to assign the covariate values of the soil association for each pipe ID in the Complete Break Database; • Categorization of pipes in the Complete Break Database into 8 groups by material and number of previous failures; • SAS statistical Software Package was used for the: • Determination of the statistical significance of variables by testing Global Null Hypothesis (Likelihood ratio, Score, and Wald); • Estimation of the baseline hazard function (time derivative of the estimated cumulative baseline hazard function, risk set equation) • Estimation of model parameters for variables; • Assessment of relative risk of failure for different criteria of soil properties; • Analysis of model results; •

Continued: Table 4-1

Reference	Model equation	Notation	variables	Methodology
(Park et al., 2010)	Proportional Hazard: Same as in (Jeffrey, 1985)	Same as in (Jeffrey, 1985)		• Grouping of 150 mm cast iron pipes into 7 ordered survival time groups according to the total number of breaks recorded; • Selection of preliminary covariates for the model (based on the criteria used for defining individual pipes); • Selection of significant covariate to be included in the model: • Univariate analysis to reveal the relationship between a covariate and the survival probability; • Examination of the statistical significance of the selected covariates using the SAS statistical system (Log-Likelihood ratio test and Akaike information); • Estimation of model coefficients of the selected variables using the phproc procedure of SAS; • Testing the proportionality assumption of the model using the hazard ratios (Score Process) • Identification of time-dependent covariate based on the results from the proportionality test; • Estimation of the baseline hazard functions: • Obtained, for each recorded failure time by using the baseline statement of the SAS system. A plot of the log-log transformed values of the baseline survival probabilities vs. the log of time indicates the appropriateness of a particular parametric model for the survival estimates. The survival function equation is obtained from the plot and the cumulative hazard function is hence calculated. The baseline hazard rates were calculated as the differences in the cumulative hazard rates between successive failure times. The baseline hazard function was then obtained by plotting a LOESS regression model to the estimated baseline hazard rates.

Table 4-2: Probabilistic multivariate models PHM and Accelerated lifetime.
Source: (Kleiner & Rajani, 2001)

References	Model	Notation	Explanatory variables (data)
Proportional hazards Marks et al. (1985)	$h(t,Z) = h_0(t)e^{b^T z}$ $h_0(t) = 2.10^{-4} - 10^{-5}t + 2.10^{-7}t^2$	T = time to next break $h(t,Z)$ = hazard function $h_0(t)$ = baseline hazard function Z = vector of covariates b = vector of coefficients to be estimated by maximum likelihood	• Natural log of pipe length • Operating pressure • Percentage of low land development • Pipe "vintage" (or period of installation) • Pipe age at second (or higher) break rate • Number of previous breaks in pipe • Soil corrosivity
Andreou et al.(1987a,b) Marks et al. (1987)	Early stage: same as Marks et al. (1985) described above Late stage: $h = \lambda = e^{b^T z}$	h = hazard (constant at the late stage)	• Same as above
Proportional hazards Brémond (1997)	$h(t,Z) = h_0(t)e^{b^T z}$ $h_0(t) = \lambda\beta(\lambda t)^{\beta-1}$	t = time to (next) break $h(t)$ = hazard function λ, β = scale and shape parameters (respectively) of the Weibull distribution	• Number of previous breaks • Pipe diameter • Ground conditions • Traffic loading
Time-dependent Poisson model Constantine and darroch (1993), Miller (1993); Constantine, Darroch, and Miller (1996)	$H(t) = \left(\frac{t}{\theta}\right)^{\beta}$ $\theta = \theta_0 e^{\propto Z}$	t = pipe age $H(t)$ = mean number of failures per unit length at age t θ, β = scale and shape parameters (respectively) θ_0 = baseline value $\propto$= vector of coefficients to be estimated by regression Z = a vector of covariates affecting breakage rate	• Mean static pressure • Overhead traffic conditions • Pipe diameter • Soil type
Accelerated life Lei (1997) ***Accelerated life*** Eisenbeis, Rostum, and Lei Gat (1999)	$\ln(T) = \mu + x^T\beta + \sigma Z$ $\Rightarrow T = f(\mu, \sigma Z)e^{x^T\beta}$ Same as Lei (1997) above	T = time to next failure x = vector of explanatory variables Z = random variable distributed as Weibull σ = parameter to be estimated by max likelihood β = vector of parameters estimated by max likelihood Z = random variable distributed as Gumbel (extreme distribution for minima)	• Pipe age group • Pipe size • Pipe length • Pipe material was taken as stratification • Log of pipe length Pipe material • Traffic loading • Soil acidity • Soil humidity Number of previous breaks was taken both as a covariate and as a stratification variate

Table 4-3: Probabilistic single-variate models.
Source: (Kleiner & Rajani, 2001)

References	Model	Notation	Data requirements
Cohort survival model Herz (1996), Deb, Hasit, Grablutz, and Herz (1998)	$f(t) = \frac{(a+1)be^{b(t-c)}}{[a+e^{b(t-c)}]^2}$ $S(t) = \frac{(a+1)}{a+e^{b(t-c)}}$ $h(t) = \frac{be^{b(t-c)}}{a+e^{b(t-c)}}$	$f(t)$ = probability density function $h(t)$ = hazard function $S(t)$ = survival function t = useful lifetime of pipe a = ageing factor ($year^{-1}$) c = resistance time (years), i.e., pipe will not be replaced at age < c years	• Pipe installation dates • Pipe "time of death" • Valid grouping criterion will enhance accuracy • Alternative to "time of death': end of economic life (optimal time for replacement) requires break history
Bayesian Diagnostic Model Kulkarni et al. (1986)	$Prob.\{failure\|specific\ characteristics\} = \frac{P_{c/f} \cdot P_f}{P_{c/f} \cdot P_f + P_{c/nf}(1-P_f)}$	P_f = system-wide probability of failure $P_{c/f}$ = probability of observing specified characteristics on a segment that has not failed	• Grouping criteria ("sets of characteristics") such as pipe diameter, length, age and type, soil characteristics, operating conditions such as pressure, etc.
Semi-Markov chain Gustafson and Clancy (1999a)	Generalized gamma distribution for t_1 Exponential distribution – identical for all $t_i (i > 1)$	t_i = time between the $(i-1)^{th}$ and the $(i)^{th}$ breaking pipe	• Pipe breakage history • Pipe type • Other grouping criteria to enhance accuracy
Break Clustering Goulter and Kazemi (1988); Goulter et al. (1993)	$P(x) = \frac{m^x e^{-m}}{x!}$ $m = m(s,t)$	m = mean number of subsequent failures occurring in the cluster domain x = number of subsequent failures occurring in the cluster domain s = distance from the 1st break in a cluster t = time elapsed from the 1st break in a cluster	• Pipe breakage history with the exact time and location of each break
Data Filtering	4 rules to filter pipe breakage data, based on calculating the probability of two consecutive breaks (Constantine and Darroch 1993), and discarding the second break if probability is low		• Pipe diameter • Pipe material • Traffic level • Soil type

4.3.9 Motivation behind the Application of the PHM in Pipe Failure Modeling

The motivation behind using the PHM in describing the time to event distribution has been repeatedly reported in the literature (see Collett, 2003; Hosmer & Lemeshow, 1999; Kalbfleisch & Prentice, 2002; Kleinbaum & Klein, 2005; Machin et al., 2006; Andreou, 1986; Jeffrey, 1985). It generally originated from the special features of survival analysis described in section (3.2.2) and particularly from the special characteristics of the Cox's PHM discussed in section4.3.1. However, variation in the model application may arise depending on each specific area of study and case study. In the following the main appealing points for the use of Cox's PHM in pipe failure modelling presented in the literature will be summarized.

One point is that the probability of break in a specific pipe is obtained directly rather than the expected number of failure in the entire system as it is the situation with the ordinary statistical regression techniques. In addition, the ability of the model to work effectively with censored data makes it possible to analyse not only pipes that fails but also pipes that have not yet experienced any failure. However, some of the recently used statistical packages (e.g. SAS and SYSTAT) cannot handle left censoring for Cox's PHM. To overcome this problem, a variable called age left can be produced which means the time from installation year to the time when failure recording starts. The model is also appealing when the focus of the analysis is to compare between survival patterns for different categories of pipes according to different explanatory variables. This can easily be done by using the stratification property of the model, (Jeffrey, 1985; Røstum, 2000).

The PHM is uniquely flexible with regard to the form of the hazard function where the baseline hazard function is unspecified and independent of the set of explanatory variables in the model. By using this property of the model, the ageing process of pipes can be independently captured through an estimation of the baseline hazard function, and so does the multiplicative effect of the included explanatory variables through the estimation of the coefficients in the exponential part of the model. This semi-parametric property of the model makes it attractive when there is doubt about appropriateness of a specific parametric model, (Kleinbaum & Klein, 2005). In general, when the objective is to evaluate the effect of covariates on the hazard function, Cox's PHM should be used. When the objective is to predict future failures within a certain time horizon; a parametric assumption about the form of the baseline hazard function is more convenient, (Kumar & Klefsjö, 1994).

One restriction on the use of survival analysis in modelling failures of water mains is that most water utilities have maintained break data for a relatively short time, which creates the problem of left censoring, (Pelletier et al., 2003). However, adjustments can be made in order to accommodate this problem. In most cases the assumption about survival times for pipes installed before recording failure incident can be made to cope with this problem. For further details see (Jeffrey, 1985; Andreou, 1986; Røstum, 2000; Park, 2004; and Park et al., 2010).

This section reviewed some of the previous survival models applied to water distribution networks. The review included a description of the model, the independent variables and the methodologies used in the model building. In the next chapter the dataset case study will be described and the various procedures of data preparation for the analysis will be discussed.

5 Case Study: The Cape Metropolitan water distribution system

5.1 Introduction

The City of Cape Town maintains data for its water distribution network which covers the area that includes the eight districts of Cape Town namely: Northern Panorama, Hillstar, Tygerberg, Helderberg, Khayelitsha, Mitchells Plain, Ebenezer, and Southern districts, (COC, 2007a), see Figure 5-1.

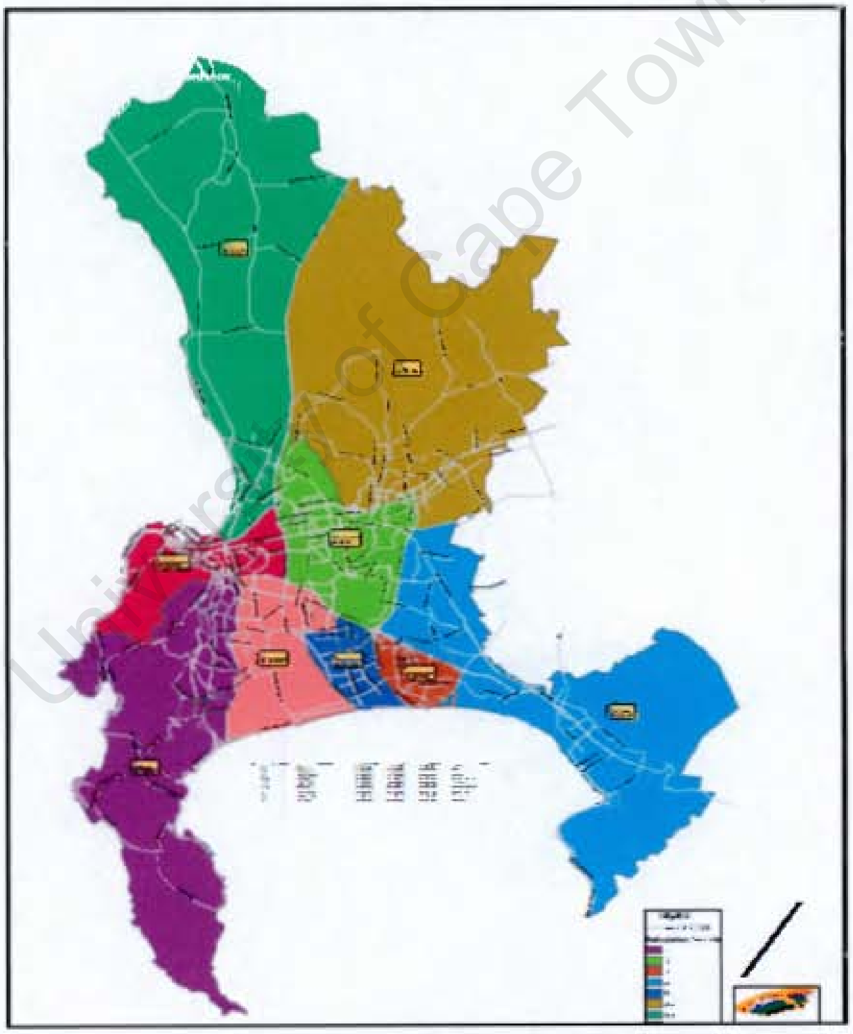


Figure 5-1: Reticulation 8 Districts and Major Roads
Source: (COC, 2007a)

According to the City of Cape Town, (COC, 2011), the total length of the water network in Cape Town amounts to about 10190 km. This is shared between different types, sizes and ages of pipes such as Fibre Cement, Cast Iron, Steel, PVC, uPVC, GRP, Prestressed concrete/steel and others. Classification of pipes based on type, size and age and their corresponding lengths are illustrated in Table 5-1 and Table 5-2.

Table 5-1: Classification of pipes based on Size and Age
Source: (COC, 2011)

Pipe size	Length (m)	Age category	length (m)
<100	1476	0 – 10	1406276
100 – 225	7149	11 – 20	3074186
250 – 450	917	21-30	1419872
500 – 675	192	31-40	1049672
675 & greater	456	41-50	1148060
Total length	10190000	51-60	266974
		over 60	1824960
		Total length	10190000

Table 5-2: Some pipe materials and their corresponding lengths in Cape Town
Source: (COC, 2011)

Pipe Material	Abbreviation	Length (km)	Percentage (%)
Prestressed Concrete	PSC	3367	0.03
Cast Iron Bitumen Lined	CIBL (bitumen)	6263	0.06
Steel Bitumen Lined	STBL	6338	0.06
Glass Fiber Reinforced Plastic	GRP	11680	0.11
Polyvinyl Chloride	PVC	30166	0.30
High Density Polyethylene	HDPE	41681	0.41
Concrete	CONCR	59802	0.59
Steel Concrete Lined	STCL	159854	1.57
Cast Iron	CI	296397	2.91
Cast Iron Concrete Lined	CICL (concrete)	495620	4.86
Steel	ST	540503	5.30
Polyvinyl Chloride (unmodified)	uPVC	558744	5.48
Fibre Cement	FC	7428679	72.90
Unknown	Unknown	550905	5.41
Total		10189999	100

In this chapter, two data sets of Cape Town's water distribution system and its failures will be used and prepared to perform statistical modeling of pipe failure in the City. First a general description of the data sets will be presented. Then the procedure used in data processing and filtering for the purpose of the general descriptive statistics and for applying the PHM to the data will be discussed. The final format of the data which will be used in the analysis and the methodology for the analysis will finally be illustrated.

5.2 Datasets Description

For the purpose of this research, two datasets were obtained from the City of Cape Town for the water distribution network. The first data set is the network data set which includes information about pipes in the network such as their location in terms of the district, the suburb and the street in which the pipes are laid, as well as the X and Y coordinates. Other information of pipes in the network includes materials, diameters, lengths and installation dates. The City also maintains data for the failures of water mains since 1980. This dataset includes information about the failing pipes such as the location of the failing pipe in terms of the suburb and the street name, the material of the pipe, the diameter, and the failure date. Further details about the two datasets will be provided in this section.

5.2.1 The Network Data

The provided network data includes different pipe materials of a total length of 2023 km. This is distributed over about 25182 pipe segments. Pipe diameters within the data vary from 20mm up to 1830 mm and their ages span from zero to 80 years or more. Classification of pipes in this dataset based on material, diameter and length is shown in Table 5-3 and Figure 5-2.

Table 5-3: Classification of network data length based on material and diameter

Diameter	<100	100-110	150-160	>160	Total (m)
Material	Length (m)				
FC	74867	599376	292844	205047	1172134
CI	1898	41353	22380	38098	103729
CICL	401	185442	66233	79546	331622
ST	-	3485	7975	172824	184284
UPVC	2735	75543	17090	1459	96827
PVC	1447	18360	1303	82	21192
HDPE	83	28613	12993	201	41890
Others	13005	16952	11078	29811	70846
	94436	969124	431896	527068	2022524

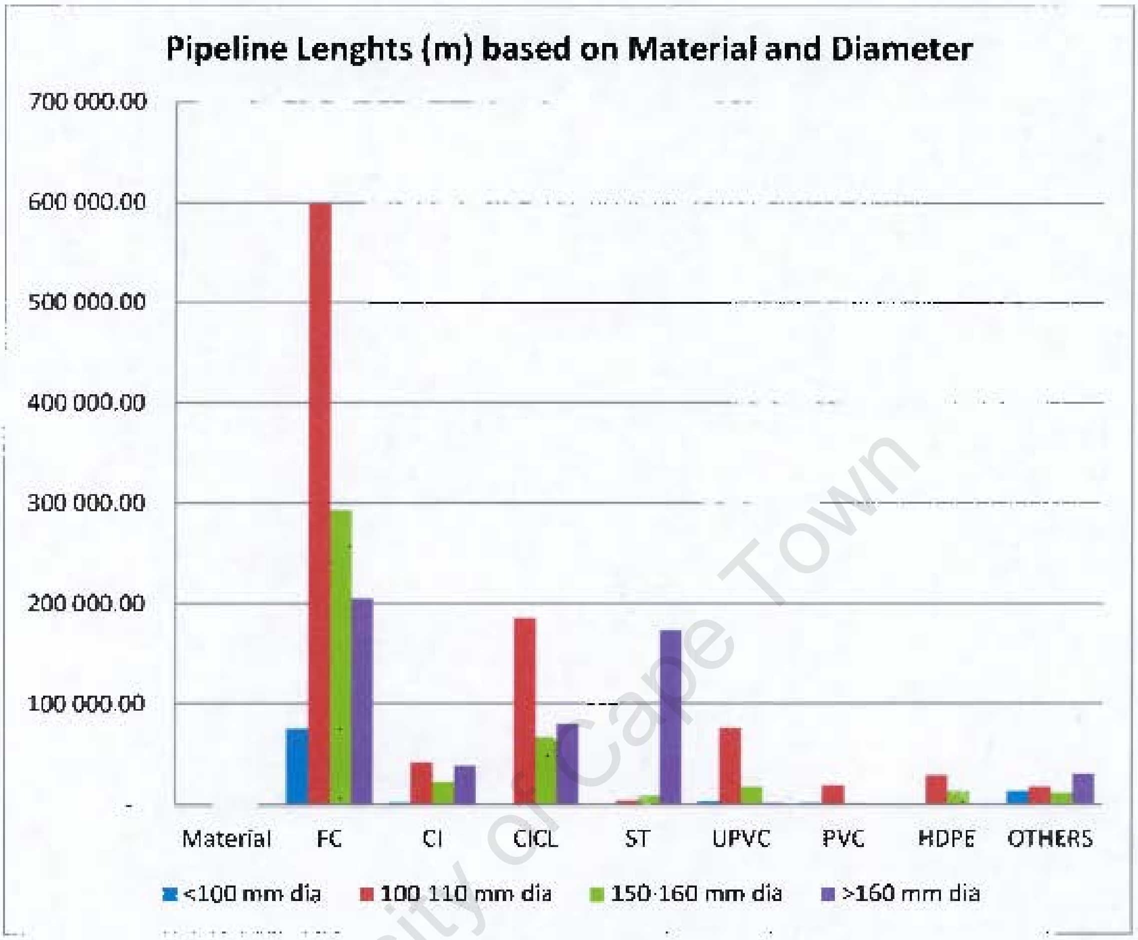


Figure 5-2: Pipeline Lengths based on Material and Diameter

Figure 5-2 shows that the (100-110) mm pipes represent the majority of the network and therefore this category has been chosen for the analysis in this dissertation and they will all be named as 100 mm FC pipes. It should also be noted that for the purpose of this work the installation dates for FC pipes have been considered to be starting since 1930.

5.2.2 The Failure Data

The total recorded number of failures in this data is 8284 failures, shared between the different pipe's materials and sizes; and spans over 30 years. Of these, 88 failures relate to 75mm pipes, 2 failures relate to 80 mm pipes and 4848 failures relate to 100 mm pipes. Details about failures records, materials and diameters are illustrated in Table 5-4 and Figure 5-3.

Table 5-4: Number of failures per material and pipe size category

Diameter category (mm)	<60	75-80	100-110	150-160	>160	Total failures per pipe type
Material	Number of failure records					
FC	315	90	4848	707	127	6087
CI	1	47	715	174	109	1046
CICL	0	1	422	71	28	522
UPVC	1	4	150	15	2	172
ST	0	0	5	9	101	115
HDPE	0	0	57	23	2	82
OTHERS	25	6	172	21	36	260
Total failures per pipe size	342	148	6369	1020	405	8284

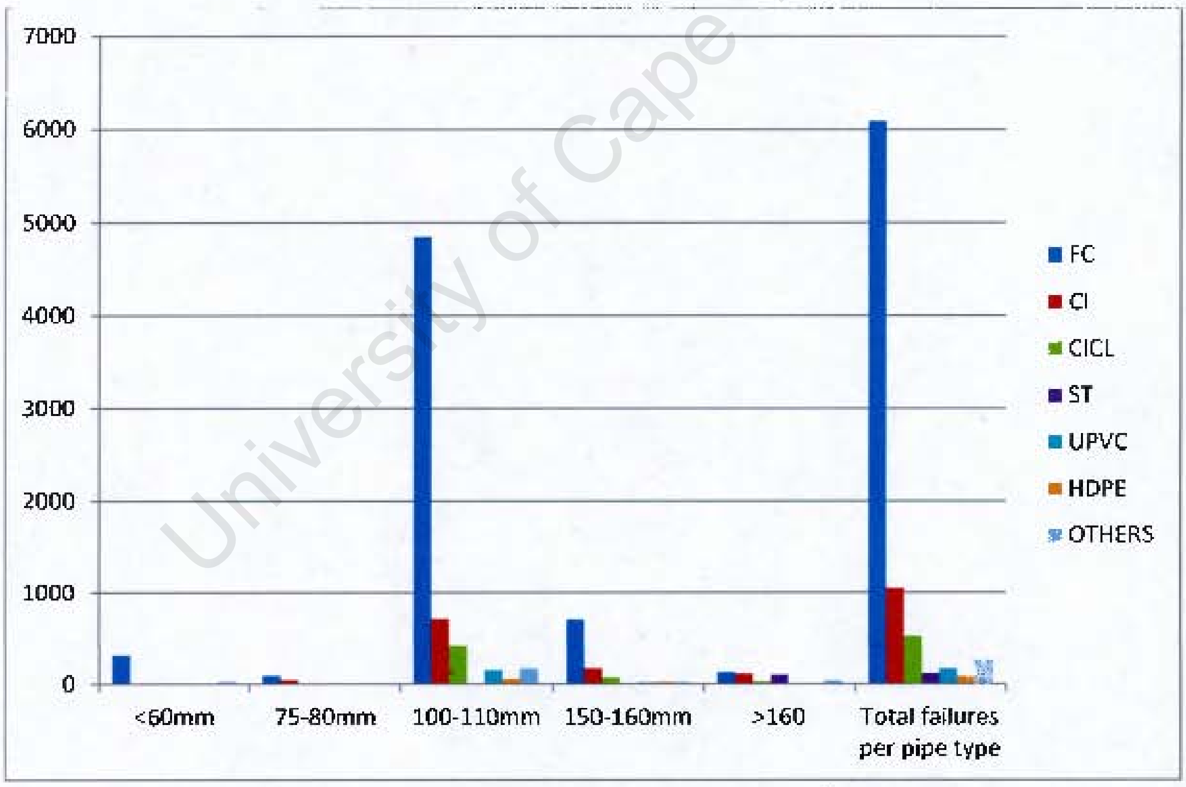


Figure 5-3: Number of failures per material and pipe size category

Figure 5-3 shows that (100-110) mm pipes encountered the highest number of failures in comparison with the other pipe types. This might be due to the fact this pipe category constitutes the majority of the network. At this stage it was impossible to display a plot of the number of the failures vs. pipe ages as they are not directly stated in any of the two datasets. This was the challenging problem, because the failure data do not reflect the age of the pipes that failed. For instance the failure data just assign the date of failure and the name of the street and suburb in which the failure incidents occurred, but it is unknown which segment of the street that relates to a specific pipe in the network and therefore the length and the installation date, which are only provided in the network data, of the failing pipe are indefinite. That makes the association between the two data sets inevitable in order to apply survival analysis which requires a certain linkage between the individual and its failure incidents and any other explanatory variables related to that specific individual. The procedure by which failure incidents have been linked to a specific pipe in the network data will be discussed in details in section 5.3.2.3.

5.3 Data Processing

It was of great importance before starting any kind of data processing to identify the type and format of data needed to perform a specific type of statistical analysis. The identification of such information depends to a great extent on the objectives of the analysis. In this thesis there were two objectives for the analysis. The first objective was to obtain a general form of the failure rate of pipes in relation with their ages and locations in terms of the suburbs in which they are laid. In order to achieve this goal a "suburb based dataset" has been developed. The first four major pipe types (i.e. FC, CI, CICL, and UPVC) has been chosen to perform this type of analysis and their ages and total lengths over the 30 years period have been calculated for each suburb.

The second objective, which is the main objective of this research, was aimed at developing a dataset for the application of survival analysis, and for this purpose the two datasets have been subjected to different cleaning, matching and programming processes. The two methods of data development are very different and will be discussed in details in this section; the assumptions used in the cleaning procedure will be highlighted and the results of the general regression analysis will also be illustrated. The description, discussion and results of the application of survival analysis will be individualized in chapter 6.

5.3.1 Datasets Preparation for the General Regression

The dependent variable in this regression is the Failure Rate (FR) defined as: number of failures/km/year. Therefore, the information needed for each pipe type in each suburb in the dataset was:

- The total number of failures in each year over the analysis period
- The total length of pipes in each year during the analysis period defined as the length of pipes of a certain age in a certain year during the analysis period.

It should be noted that in this research, and due to some mistakes in the provided datasets, some corrections and assumptions have been made to the dataset and these are:

- The installation date for some pipe segments with unknown or invalid installation dates (e.g. 1900) is considered to be the same as the other surrounding pipes laid in the same street with typical or consistent installation dates.
- Due to the fact that CI pipes have all been lined between 1969 and 1980, they have been joined to the CICL pipes; and hence all the CI pipes in one street with unknown or invalid installation dates have been considered to be installed in 1969.
- Due to the fact that FC pipes have only been put into service in Cape Town in the late 1930s, all the FC pipes with unknown or invalid installation dates have been considered to be installed in 1930.
- The installation of UPVC has generally been assumed to be a replacement of a CICL or FC pipe (unless it was a new development). The presence of a UPVC pipe in 2010 has, therefore, interpreted as that at some earlier time there was the same length of either FC or CICL pipe and so the lengths of the UPVC pipes have been adjusted based on this assumption and also PVC pipes have been joined with UPVC ones, (DelMistro, 2011).

The methodology of data preparation and the results of the analysis are discussed in the next section.

5.3.1.1 Dataset Development for the General Regression

In this approach the analysis period (i.e. the 30 years) has been categorized into two years span and hence considered as the cases of the regression. The independent variables, also known as the regressors, have been calculated for each suburb and defined as: **The length of pipes of a certain age category during a certain analysis period category.**

The basis for the calculation of these lengths was the total length of pipes in 2010. For instance if the total length pipes of age 80 years in 2010 was 2243, it is expected that this would be the same length of pipes of age 78 years in 2008 and so on. These lengths have been adjusted to accommodate the presence of UPVC pipes in 2010. For this purpose the pipes have also been grouped into two years span of age, and the total length of pipes that correspond to a specific age and analysis period category has been calculated. An example of the format of the data at this stage for some suburbs is shown in Table 5-5.

On the failure data side, the average number of failures that correspond to these pipe lengths has been calculated for the same two years span of analysis for each suburb. These lengths have then been aggregated to include the entire network and the Failure Rate, considered as the independent variable, have been obtained by dividing the total average numbers of failures for all suburbs over the analysis period by the total lengths of pipes in all suburbs over the 30 years period of analysis.

Many investigatory regression runs have firstly been performed on the 100 mm FC pipes with different independent variables in order to obtain the best results for the t-stat, regression coefficients and the adjusted R^2. The regressions have been undertaken using the MS Excel program and the final format of the dataset is shown in Table 5-6.

The dataset has been formatted for the first regression run (i.e. Regression 1) by deleting all the blank lines, columns of ages without pipe lengths of that age and also some suburbs without failure data. In this run the resultant number of variables was 40, however, and due to some limitations related to the number of variables in Excel as it can only analyse up to 16 variables, 2 columns in every 3 have been deleted. The second regression run (Regression 2) kept the format of the first run but the failures of 2010 have been multiplied by 2 to account for the fact the data have only been obtained for the first half of 2010. Instead of multiplying the failures of 2010 in (Regression 2 by 2), this year has been entirely deleted from the analysis in (Regression 3). It has been found from (Regressions 3) that age 78 produces a very poor t-stat; hence it has been deleted to result in (Regression 4). In Regressions (5) and (6), the data for 1980 and 1982 were deleted because there were no/or very few failures for them.

A scatter plot and a line graph of the coefficients from Regression run (6) have been made and shown in Figure 5-4 and Figure 5-5.

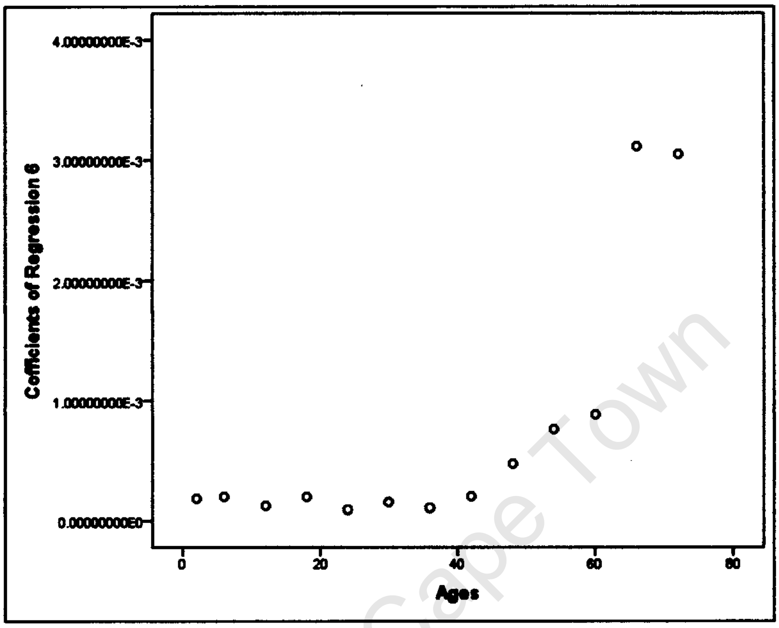


Figure 5-4: Scatter plot of regression coefficients (6) vs. Age

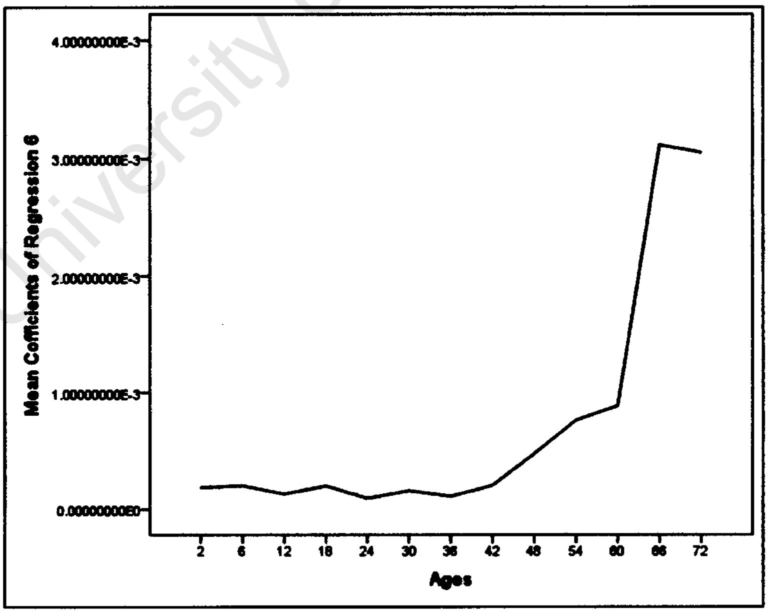


Figure 5-5: Line graph of regression coefficients (6) vs. Age

The scatter plot and the line graph of regression coefficients from run (6) suggests that the failure rates for FC pipes have the bathtub curve which distinguish between three stages of pipe's age. The first stage represents pipes up to age 42 years which represents stable failures. Pipes are then experienced normal increase in the failure rates from 42 – 60 years which represent the start of the wear out stage of the bathtub curve followed by dramatic increase in failure rates for older pipes.

Although the approach resulted in acceptable values for the regression coefficients and t-stats, the adjusted R^2 didn't seem to be so good, (a summary table of the results of the different regression runs is shown in Table 5-7). Correlation tests have been performed and the results showed that most of the independent variables are highly correlated. This might be attributed to the dependency of the calculation of the total length for the different age categories on the total length of pipes in 2010 and the huge amount of aggregation in the calculation of the failure rate. To this end, this approach of data format has been stopped and a new methodology for the dataset development has been proposed for the general regression.

Table 5-5: Dataset preparation for the general regression (1st approach)

	AGE	108	106	104	102	100	98	96	94	92	90	88	86	84	82	80	78	76	74	72
suburbs	Total pipe length by age and analysis period category																			
Athlone	2010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2243	0	0	0
	2008		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2243	0	0
	2006			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2243	0
	2004				0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2243
	2002					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2000						0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1998							0	0	0	0	0	0	0	0	0	0	0	0	0
	1996								0	0	0	0	0	0	0	0	0	0	0	0
	1994									0	0	0	0	0	0	0	0	0	0	0
	1992										0	0	0	0	0	0	0	0	0	0
	1990											0	0	0	0	0	0	0	0	0
	1988												0	0	0	0	0	0	0	0
	1986													0	0	0	0	0	0	0
	1984														0	0	0	0	0	0
	1982															0	0	0	0	0
	1980																0	0	0	0
Bantry Bay	2010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2008		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2006			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2004				0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5-6: Final format of data for the general regression 100 mm FC (1st approach)

Age	78	72	66	60	54	48	42	36	30	24	18	12	6	2	Average # Failures/year	failure/km/yr.
Analysis period categories	Total length of pipes for all suburbs by age and analysis period category															
2008	0	286	5488	2307	2312	19350	33990	14281	8463	35640	26507	5267	5674	0	133	0.41519
2006	0	0	1698	19843	1930	8175	11665	7552	22748	29529	22713	16287	1419	0	147	0.51199
2004	0	19318	455	3477	151	1430	49743	26433	40577	12715	39161	36900	1243	5674	133	0.28026
2002	0	0	286	5488	2307	2312	19350	33998	14500	8463	35640	26569	5267	1419	135	0.43220
2000	0	0	0	2051	20436	2105	8458	12111	8042	23068	30777	24398	17316	1280	121	0.40322
1998	0	0	23609	515	3748	151	1670	50878	28459	42294	13817	41472	41016	6280	120	0.23532
1996	0	0	0	307	6183	2631	2823	29565	37321	16735	9446	45296	28534	18307	105	0.26503
1994	0	0	0	0	2406	21196	2328	8884	13427	9286	23972	32767	25706	42210	112	0.30738
1992	0	0	0	27558	597	3895	173	1782	51323	29063	44990	14759	42336	28678	111	0.22537
1990	0	0	0	0	310	6299	2656	2900	29800	38039	16913	9554	45566	25748	92	0.25734
1988	0	0	0	0	0	2406	21196	2329	8886	13586	9342	25360	32883	42653	70	0.22062
1986	0	0	0	0	27650	597	3895	173	1785	51356	29064	44990	14863	46021	76	0.17128
1984	0	0	0	0	0	310	6299	2656	2900	29803	38078	16936	9556	33047	57	0.20239
1982	0	0	0	0	0	0	2406	21200	2332	8886	13973	9370	25688	14863	9	0.04558
1980	2212	0	0	0	0	28071	604	3903	173	1785	51497	29192	45815	9820	0	0.00000

Table 5-7: A summary of the general regression trials results (1st approach)

	(Regression 1): Blank lines deleted & sets without failure data; i.e Salt River& Three Anchor Bay.		(Regression 2): Blank lines deleted & sets without failure data; half year 2010 failures * 2		(Regression 3): Blank lines deleted & sets without failure data; delete 2010 (3)		(Regression 4): Blank lines deleted & sets without failure data; delete 2010, delete age 78		(Regression 5): Blank lines deleted & sets without failure data; delete 2010; delete 1980 because failures=0 (		(Regression 6): Blank lines delete & sets without failure data; delete 2010;/ delete 1980 and 1982 - no and low failures (6)	
R Square	0.212		0.227		0.216		0.216		0.224		0.226	
Adjusted R Square	0.200		0.215		0.204		0.205		0.212		0.212	
Observations	953		953		893		893		835		777	
	df	F	df	F	df	F	df	F	df	F	df	F
Regression	14	18.05	14	19.66	14	17.35	13	18.70	13	18.30	13	17.13
	Coefficients	t Stat	Coefficients	t Stat	Coefficients	t Stat	Coefficients	t Stat	Coefficients	t Stat	Coefficients	t Stat
78	0.00098	1.76	0.00246	4.36	-0.00020	-0.10						
72	0.00310	5.37	0.00307	5.25	0.00307	5.18	0.00307	5.19	0.00305	5.00	0.00305	4.83
66	0.00289	7.49	0.00293	7.48	0.00315	7.56	0.00315	7.56	0.00313	7.29	0.00312	7.00
60	0.00092	5.12	0.00092	5.04	0.00090	4.88	0.00090	4.89	0.00089	4.67	0.00089	4.50
54	0.00078	4.45	0.00077	4.37	0.00077	4.31	0.00077	4.31	0.00076	4.15	0.00076	4.00
48	0.00009	1.36	0.00009	1.34	0.00038	2.64	0.00038	2.65	0.00048	2.97	0.00047	2.86
42	0.00020	3.26	0.00021	3.39	0.00021	3.18	0.00021	3.18	0.00020	3.03	0.00020	2.92
36	0.00012	2.37	0.00015	2.84	0.00010	1.81	0.00010	1.81	0.00010	1.71	0.00011	1.72
30	0.00016	3.15	0.00017	3.13	0.00016	3.04	0.00016	3.05	0.00016	2.88	0.00016	2.76
24	0.00009	1.84	0.00009	1.88	0.00010	1.93	0.00010	1.93	0.00009	1.81	0.00009	1.74
18	0.00011	2.29	0.00011	2.25	0.00011	2.28	0.00011	2.28	0.00020	2.99	0.00020	2.91
12	0.00013	2.12	0.00012	2.06	0.00012	2.00	0.00012	2.00	0.00013	1.95	0.00013	1.85
6	0.00015	2.31	0.00014	2.25	0.00014	2.20	0.00014	2.21	0.00018	2.36	0.00020	2.42
2	0.00020	2.64	0.00020	2.59	0.00020	2.54	0.00020	2.54	0.00019	2.31	0.00019	2.22

5.3.2 Datasets Preparation for the Application of Survival Analysis

A major assumption when applying survival analysis to a repairable system, as is the case for water distribution systems is that the pipe section is considered as a new pipe section after repair (Park et al., 2008; Park et al., 2010). Each pipe will be considered as a new pipe after each failure record and therefore there will be as many survival times for one pipe as its associated number of failure records. It is therefore essential, in order to identify the survival times for individual pipes, for the dataset to have all the installation dates and failure incidents of pipes in the network identified and associated.

As has been mentioned earlier in this chapter that a major problem with the datasets provided for this research is that the network data and the failure data are not fully linked to each other, therefore, the failure records are not assigned to specific pipes in the network. To overcome this problem and have the two datasets linked to each other; and to extract the desirable information from them; they have been subjected to an extensive matching, filtering and programming processes. These processes should, however, be governed by certain definitions and criteria by which failure incidents and individual pipes should be assigned.

This section discusses the varies stages, definitions, and assumptions by which the 100 mm FC pipes in the network dataset and the failure dataset have been prepared for the application of the two approaches of survival analysis.

5.3.2.1 Definition of Failure Incidents

In this research pipeline failure is defined as failure of a pipe recorded as a repair activity that required municipal intervention.

5.3.2.2 Definition of Individual Pipes

The identification of individual pipes is a very important and critical part of the analysis. In this research a group of pipe segments whose installation dates do not vary by more than 15 years and that are laid in one street is defined as an individual pipe. The installation date of the new pipe is defined as the average of the installation dates of the original pipe segments and the length of the new pipe is defined as the sum of their lengths.

5.3.2.3 Fitting Failure Records to Actual Pipes

In order to develop a survival dataset, failure records in the failure dataset should be associated with the specific related pipes in the network data so that all the information can be found in one dataset. The link between the two datasets which can be used for such association is the street names and suburbs. Other information in one of the two dataset could not be used as it doesn't exist in the other dataset.

Furthermore, there was a lack of standardization and similarity between the two datasets in recording street names and suburbs, therefore an automated association between the two dataset was impossible at this stage. For these reasons manual association between the two datasets was found to be the appropriate method.

4848 Failure records for the 100 mm FC pipes have been individually checked for potential association with their related pipe segments in the network dataset using the GIS system maps (Google Earth). The main target of this process of association is to obtain a dataset with a great degree of accuracy with regard to the relation between the individuals (i.e. the pipe) and its explanatory variables (i.e. age, diameter, length, number of previous failures, etc.) so as to develop the most accurate and reliable model possible for failure time.

To this end the only streets that are considered in the analysis are the streets which can be directly and accurately linked. Other failure records which need further assumptions regarding the location or the age of the pipe segments in order to be associated are not included in the analysis. In this first stage of failure records – pipes segments association, 2175 failure records out of 4848 failure records have been fully and accurately linked with a total of 2271 of 7251 pipe segments in the network data. The percentage of the matched failure data is shown in Figure 5-6. The resultant data sets are shown in Table 5-10 and Table 5-11.

Having failure records associated with certain pipe segments, the next step was to combine the two datasets in one dataset that, recalling the definition of the individual pipe introduced in section 5.3.2.2; aggregates pipe segments of one street into one individual pipe and to relate the matched failure records to that assigned street. A computer program has been developed for this purpose using the MS Access program and the resultant number of individuals was 540. The output format of the data is shown in Table 5-12. Of these 540 individual pipes, 35 resulted in negative age to failure (i.e. first failure date –

installation date = negative value) and therefore the final total number of individuals that will be included in the analysis is 505 individual pipes.

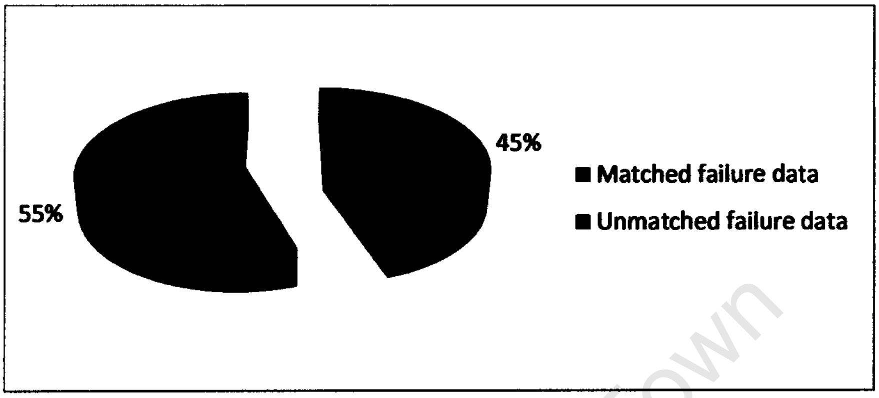


Figure 5-6: Percentages of matched and unmatched failure records

5.3.2.4 Primary Selected Variables in the Dataset

The primary selected variables have been formatted using the output dataset shown in Table 5-13 as

- Survival times or Age (installation-failure): for each individual pipe it corresponds to the age of the pipe at the time of failure. (i.e. time from installation to the first failure and time from installation to the second failure) given by

  Age (installation- failure1) = failure date – installation date

  recalling the discussion held in section 3.2.1, the survival time of a pipe may differ according to the defined time origin of the study. For instance in this research, the time origin has been considered as from the date in which the first pipe was laid, i.e. 1930. In other case it could be considered as the time of the start of the study, i.e. 1980.The survival time for the case study dataset has been calculated for each pipe for each failure occurrence. Pipes that failed at their year of installation will have a zero survival time, Others that experienced two failures in the same year will have the same survival time for the two failures and those with no survival time, (i.e. haven't experienced a failure or censored) are denoted not applicable (N/A).

- Length: corresponds to the total length of the defined individual pipe. The length of an individual pipe is assumed to remain the same for each failure occurrence. In many application settings this assumption is not correct as it ignores the extension of the water network over time; however, it might be justified for the case study dataset due to the lack of standardization between the network dataset and the failure dataset. Pipe lengths have been classified into different groups during the analysis to suit a special statistical setting. These will be illustrated for each case as needed.

- Number of previous failures: this variable has been prepared by counting the total number of failures over the analysis period for each individual pipe. For different purposes of analysis, pipes have been grouped into four categories according to the number of previous failures as shown in Table 5-8.

Table 5-8: Stratification of pipes according to the NOPF

Number of previous failures	Code
1-4	1
5-10	2
11-20	3
>20	4

- Installation era: this variable has been included to accommodate the possibility of manufacturing defects in a specific installation periods and used in the analysis as a dummy variable. Individual pipes have been classified according to their installation periods into 11 eras as shown in Table 5-9

Table 5-9: Installation eras and their codings

Installation Era	Code
1900-1910	1
1911-1920	2
1921-1930	3
1931-1940	4
1941-1950	5
1951-1960	6
1961-1970	7
1971-1980	8
1981-1990	9
1991-2000	10
2001-2010	11

- Censorship: the variable censorship denotes the status of the individual pipe after each failure and it is entered to the dataset as a dummy variable. If a pipe failed in a specific date period and this failure is known to be the last failure for this pipe (i.e. there is no subsequent recorded failure) then it is labeled "0." i.e. the pipe is censored after the last failure. Conversely, a label "1" is given to a pipe that is known to have failed in a specific date and this failure is known to have a subsequent recorded failure.

The final data set prepared for the application of survival analysis that includes the basic variables is illustrated in Table 5-14. Other transformations made to variables in the dataset during the analysis for both approaches will be described as needed.

Table 5-10: Matched network data sample

Street Name	Numeric Installation Date	Corrected Installation Date	Material	Length	Diameter
Active Road	1971	1971	FC	111	100
Adrian Herbert Road	1973	1973	FC	200	100
Allies Road	1900	1930	FC	362	100
Allison Road	1963	1963	FC	10	100
Allison Road	1964	1964	FC	24	100
Allison Road	1965	1965	FC	68	100
Allison Road	1969	1969	FC	143	100
Allison Road	1969	1969	FC	87	110
Allison Road	1969	1969	FC	88	110
Almar Road	1941	1941	FC	186	100
Almar Road	1941	1941	FC	164	100
Almond Terrace	1983	1983	FC	37	100
Aloe Street	2001	2001	FC	68	100
Aloe Street	2001	2001	FC	336	100
Amazon Street	1900	1969	FC	166	100
Amazon Street	1969	1969	FC	442	100
Amity Way	1959	1959	FC	160	100
Anadale Road	1994	1994	FC	342	100
Anglesey Street	1977	1977	FC	2	100
Anglesey Street	1977	1977	FC	85	100
Anglesey Street	1977	1977	FC	43	100
Anglesey Street	1977	1977	FC	34	100
Anglesey Street	1977	1977	FC	36	100
Anglesey Street	1977	1977	FC	6	100
Antelope Court	1956	1956	FC	64	100
...	...	...	...	...	...

Table 5-11: Matched failure data sample

Street Name	Suburb	Material	Diameter	Failure Date
Active Road	Penlyn Estate	FC	100	1993
Adrian Herbert Road	Wetton	FC	100	2002
Adrian Herbert Road	Wetton	FC	100	2002
Adrian Herbert Road	Wetton	FC	100	2005
Adrian Herbert Road	Wetton	FC	100	2005
Adrian Herbert Road	Wetton	FC	100	2006
Allies Road	Newfields	FC	100	1983
Allison Road	Pinati	FC	100	1998
Allison Road	Pinati	FC	100	2000
Allison Road	Pinati	FC	100	2001
Allison Road	Pinati	FC	100	2002
Allison Road	Pinati	FC	100	2003
Allison Road	Pinati	FC	100	2004
Allison Road	Pinati	FC	100	2005
Almar Road	Lansdowne	FC	100	1992
Almar Road	Lansdowne	FC	100	1998
Almar Road	Lansdowne	FC	100	1998
Almar Road	Lansdowne	FC	100	2006
Almond Terrace	Schotschekloof	FC	100	2004
Aloe Street	Hanover Park	FC	100	2001
Amazon Street	Primrose Park	FC	100	1985
Amazon Street	Primrose Park	FC	100	1990
Amazon Street	Primrose Park	FC	100	1994
Amazon Street	Primrose Park	FC	100	1998
Amazon Street	Primrose Park	FC	100	1998
Amazon Street	Primrose Park	FC	100	2000
Amity Way	Lansdowne	FC	100	1988
Amity Way	Lansdowne	FC	100	1990
...	...	...	...	...

Table 5-12: Output data from the MS Access Program

Street name	Average of corrected installation date	Sum of length	Failure No 1	Failure No 2	Failure No 3	Failure No 4	...
Active Road	1971	111	1993				...
Adrian Herbert Road	1973	200	2002	2002	2005	2005	...
Allies Road	1930	362	1983				...
Allison Road	1967	420	1998	2000	2001	2002	...
Almar Road	1941	350	1992	1998	1998	2006	...
Almond Terrace	1983	37	2004				...
Aloe Street	2001	405	2001				...
Amazon Street	1950	608	1985	1990	1994	1998	...
Amity Way	1959	160	1988	1990	1994	1996	...
Anadale Road	1994	342	1990	1991	1991	1993	...
Anglesey Street	1977	207	2007	2009			...
Antelope Court	1956	128	2001				...
Arcadia Avenue	1976	47	1991	2003	2006	2009	...
Arlington Road	1968	235	1995	1995	1997	1998	...
Athall Walk	1972	559	2002	2005			...
...	...	...	...	...	...	...	...

Table 5-13: Primary selected variables for the application of survival analyses

Street Name	Average of installation date	Sum Of Length	No. of Prev. Failures	Age1(Inst-failure1)	Sum Of Length	NOPF	Installation era	Censorship1(1 if failed, (0)if censored)	Age2(inst-failure2)	Sum Of Length	NOPF	...
Active Road	1971	111	1	22	111	1	8	0	N/A	111	1	...
Adrian Herbert Road	1973	200	5	29	200	5	8	1	29	200	5	...
Allies Road	1930	362	1	53	362	1	3	0	N/A	362	1	...
Allison Road	1967	420	7	32	420	7	7	1	34	420	7	...
Almar Road	1941	350	4	51	350	4	5	1	57	350	4	...
Almond Terrace	1983	37	1	21	37	1	9	0	N/A	37	1	...
Aloe Street	2001	405	1	0	405	1	11	0	N/A	405	1	...
Amazon Street	1969	608	6	16	608	6	7	1	21	608	6	...
Amity Way	1959	160	9	29	160	9	6	1	31	160	9	...
Anglesey Street	1977	207	2	30	207	2	8	1	32	207	2	...
Antelope Court	1956	128	1	45	128	1	6	0	N/A	128	1	...
Arcadia Avenue	1976	47	4	15	47	4	8	1	27	47	4	...
Arlington Road	1968	235	10	27	235	10	7	1	27	235	10	...
Athall Walk	1972	559	2	30	559	2	8	1	33	559	2	...
Athburg Walk	1972	965	27	16	965	27	8	1	20	965	27	...
Athon Walk	1975	632	4	18	632	4	8	1	26	632	4	...
Athsur Road	1972	372	4	13	372	4	8	1	25	372	4	...
Athwood Road	1975	797	7	17	797	7	8	1	17	797	7	...
Avenger	1980	83	1	28	83	1	8	0	N/A	83	1	...
...	...	...	...	...	...	...	...	...	...	...	...	...

5.3.2.5 Data Set Development for the Estimation of the MCF

This section discusses the various stages of the development of the dataset for the estimation of the Mean Cumulative Function (MCF) from exact data with right censoring. The dataset shown in Table 5-13 has been used and formatted based on the methodology described by (Nelson, 2003) as

1. **Survival times:** Each failure recurrence and censoring ages for the 505 individual pipes (shown in Table 5-13) have been ordered from smallest to largest, as shown in column 1 of Table 5-14. These amounts to 1762 survival times. Censoring status that corresponds to each sample recurrence has been denoted in another column by putting (0) if the age is censored and (1) if it is failed.
2. **Length of pipes at risk:** For each sample age, write in column 6 the length of remaining units ("at risk" of recurrence) L at that age as follows. For the 5th age for instance, write the total remaining length of pipes from the 4th age minus the length of the 5th pipe if it is censored. Otherwise, write the total length of pipes if it is a recurrence (i.e. the same length for the pipes from the 4th age). Then proceed down the column writing the preceding length for each recurrence age, or writing the preceding length minus the length of the pipe of that age for each censoring age. That is, the remaining length of pipes decreases at each censoring age by the length of the censored pipe at that age. For the last age, which is always censored, the length = 0.
3. **Mean number:** The observed incremental "mean number of recurrences per unit" at that age has been calculated in column 7 for each recurrence as 1/L. That is, one out of the length that passed through that age. For a censoring age, the observed mean number of recurrences will not be included. The censoring ages determine the length values of the recurrences; thus the calculation takes into account the censoring ages.
4. **MCF:** In column 4, calculate the value of the sample MCF $M(t)$ at each recurrence by summing the preceding increments as follows. For the earliest recurrence age, its MCF value is its mean number of recurrences in column 7. For each successive recurrence age, its MCF value is its incremental mean number of recurrences (column 7) plus the preceding MCF value (column 8). No MCF value is calculated for censoring ages, but they are taken into account, as they determine the number at risk for each recurrence. The final dataset for this approach is shown in Table 5-14.
5. **Plot:** The horizontal age scale has been chosen to include the range of the age data, and the vertical MCF scale has been chosen to include the MCF values. On the graph, plot

each recurrence's MCF value (column 8) against its age (column 1). Censoring ages are not plotted. This plot displays the nonparametric estimate $M(t)$, also called the sample MCF. The sample MCF extends only to the last censoring age. Therefore, it is useful to display that age (and all other censoring ages) on the age axis to show the range of the data.

Table 5-14: Final format of dataset for survival analyses

Age (installation - failure)	Length (Km)	NOPF	Installation Era	Censorship (1 failed, 0 censored)	Total Length at Risk (L)	Mean number/ L	MCF
1	2	3	4	5	6	7	8
					[redacted]		
0	399	1	9	0	170092	0.0000058792	0.0000000000
0	405	1	11	0	169687	0.0000058932	0.0000000000
0	337	1	10	0	169351	0.0000059049	0.0000000000
0	76	2	9	1	169351	0.0000059049	0.0000059049
0	429	1	9	0	168922	0.0000059199	0.0000059049
0	6	2	9	1	168922	0.0000059199	0.0000118248
0	166	2	9	1	168922	0.0000059199	0.0000177447
0	300	5	10	1	168922	0.0000059199	0.0000236646
1	387	1	9	0	168535	0.0000059335	0.0000236646
1	167	1	9	0	168368	0.0000059394	0.0000236646
1	158	4	10	1	168368	0.0000059394	0.0000296040
1	166	2	9	0	168202	0.0000059452	0.0000296040
1	240	12	9	1	168202	0.0000059452	0.0000355492
1	161	1	9	0	168041	0.0000059509	0.0000355492
2	338	1	9	0	167703	0.0000059629	0.0000355492
2	88	1	9	0	167615	0.0000059660	0.0000355492
2	601	1	9	0	167015	0.0000059875	0.0000355492
2	158	4	10	1	167015	0.0000059875	0.0000415367
...	...	...	...	...	...	...	...

5.3.2.6 Data Set Development for the Application of the PHM

The dataset shown in Table 5-14 have been used for the application of the PHM in the case study. However the identification of significant variables for the model is constructed through extensive statistical procedures through which the final format of the model's data can be identified.

The methodology used for the application of the PHM in this case study is based on the methodologies adopted by (Jeffrey, 1985; Andreou, 1986; Park, 2004; and Park, et al, 2010) and on the description of the model application in (Kleinbaum & Klein, 2005 and Machin et al., 2006). The methodology is illustrated in details in the schematic diagram of Figure 5-7.

The first step in their methodologies was to identify the range and variability of the variables in the dataset (e.g. survival time, length, diamether, number of previous failures, installation eras etc.). This has already been done and prepared for the case study in Table 5-14. The next step is the primary selection of explanatory variables for the model. This can be done by performing a number of univariate and bivariate analysis for the number of previous failure aginst some of the explanatory variables. Then a correlation analysis should be undertaken to examine the statistical significance of these variables.

Having selected the initial explanatory variables, it is time to perform survival analysis to further examine the statistical significane of these variables and to identify other factors where appropriate. For this purpose pipes have been grouped according to the number of previous failures and a PHM is applied in order to examine the statistical significance of the included factors. Pipes are then stratified into different groups according to the initially selected variable and a PHM is then performed for each stratum in order to identify different failure patterns within and between groups of pipes, and to examine the statistical significance of these variables. It should be noted that all the above mentioned statistical procedures can be made using the statistical software SPSS.

Depending on the results from the preformed survival analysis, a PHM is then applied to the whole dataset, the regression coefficients are determined and the final model at pipe level is configured as well as at stratum level if desired. The application of this methodology in the case study will be discussed and analyzed in chapter 6.

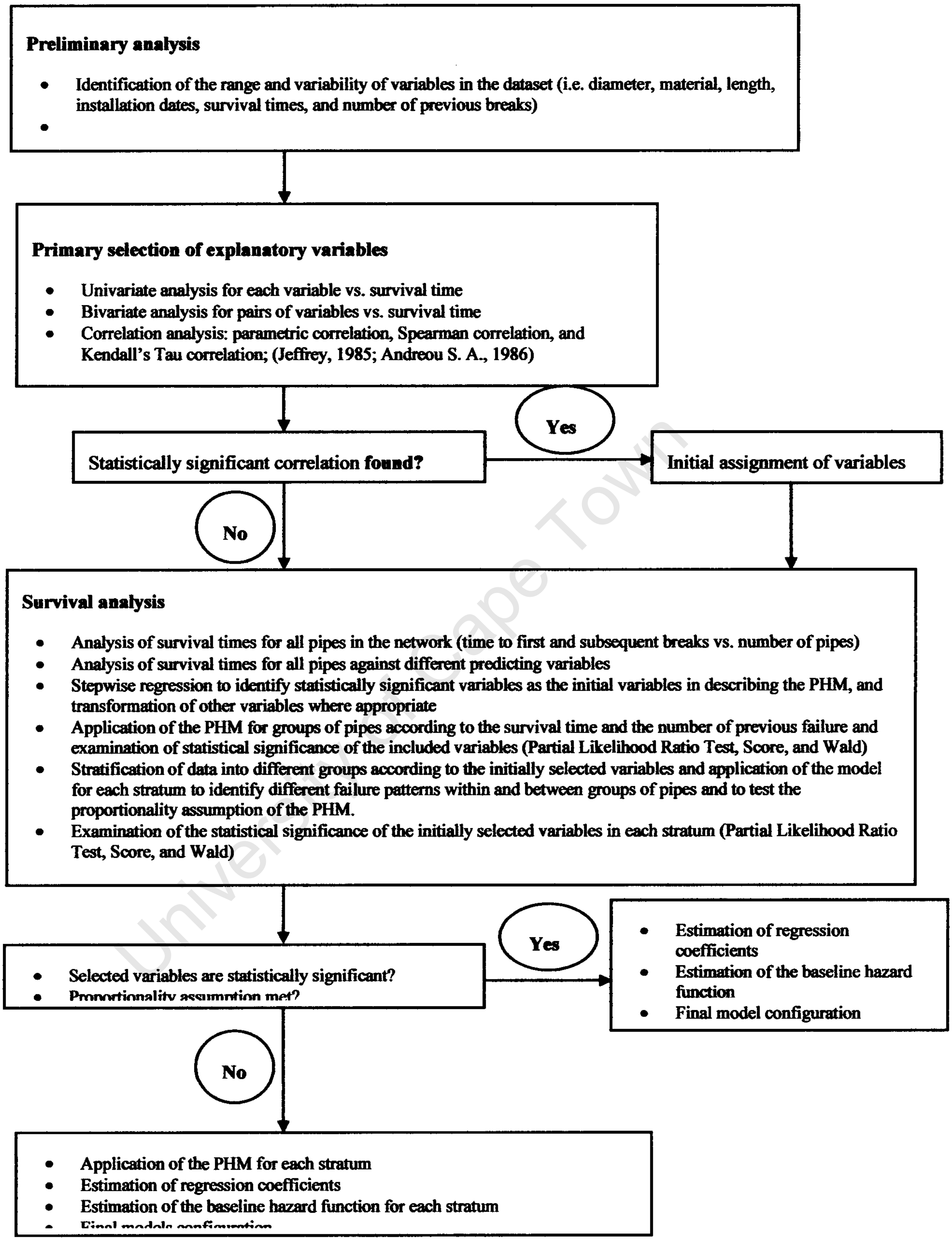


Figure 5-7: Proposed methodology for the application of the PHM

6 Analyses and Results

In this chapter, the dataset shown in Table 5-14 have been used to develop statistical models that describe the Failure Rate (FR) of the 100 mm FC pipes as a function of the pipe age, and to assess the effect of the different factors influencing the failure by applying the PHM to the dataset.

6.1 MCF Estimates VS Age

For the purpose of the development of failure rate models, MCF estimates (i.e. column 8 in Table 5-14) have been plotted against survival ages (i.e. column 1 in the same table) using the SPSS statistical package as described in section 4.2.1. A scatter plot of the two variables is shown in Figure 6-1.

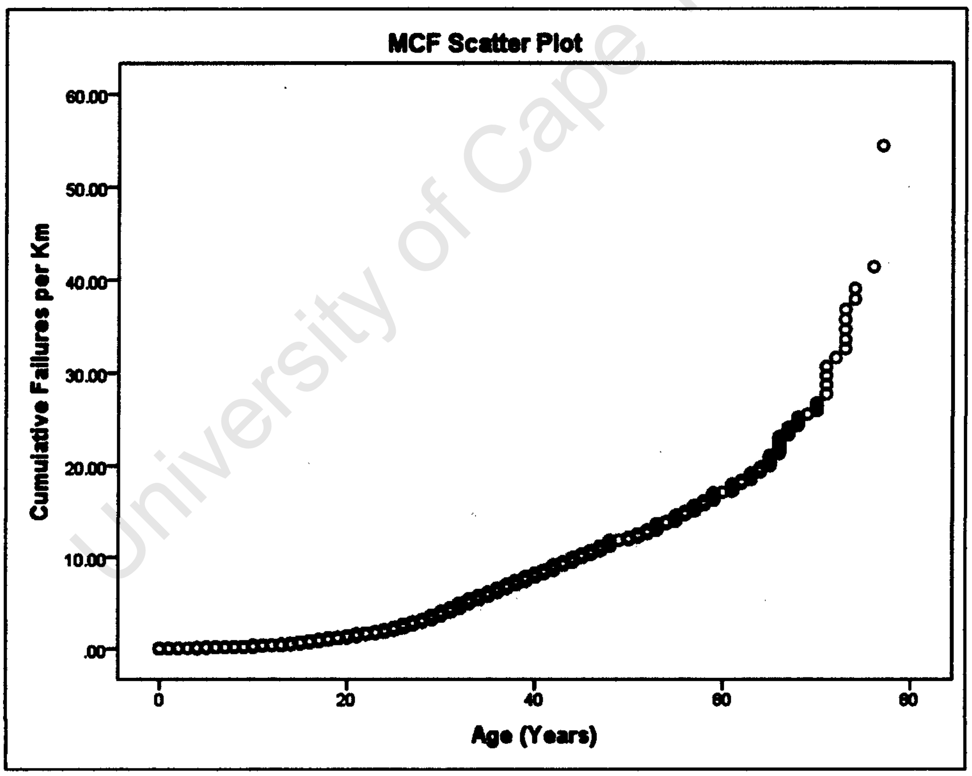


Figure 6-1: MCF scatter plot with Age based on Nelson's nonparametric estimate

As expected the plot implies an increased failure rate (i.e. the slope of the curve) with age. It typically depicts the bathtub curve shape. Approximately, one can distinguish between three stages of failure rate (i.e. the slope of the MCF curve):

- Stage 1: for ages between 0 – 25 years, the failure rate is relatively stable with very few cumulative failures per km until the age 25 years (about 3f/km) and almost negligible for ages less than 15 years.
- Stage 2: for ages between 26 – 65 years, the failure rates have smoothly increased and the cumulative failures have reached about 20f/km until age 65 years
- Stage 3: for ages greater than 65 years the failure rate has steeply increased to with corresponding cumulative failures that amount to more than 55f/km until age 78 years

Stage one represents pipes that are in the in-service or stabilized failure rate whereas stage 2 represents pipes that are at the beginning of the wear out stage and stage 3 represents pipes that are actually wearing out.

The failure rate which is the main target of this estimation can be obtained by first fitting the MCF estimates to a suitable curve and then taking the time derivative of the estimated curve. The number of failure at certain age can simply be found by multiplying the MCF by the length of pipe of concern.

6.1.1 MCF Stratification by Number of Previous Failures

To examine any trends for failure rate of pipes in relation to the number of previous failures, the MCF estimates have been plotted against the survival times according to the classification of the number of previous failures of pipes defined in Table 5-8. No noticeable trends of the failure were portrayed for the different categories of the number of previous failures, see Figure 6-2.

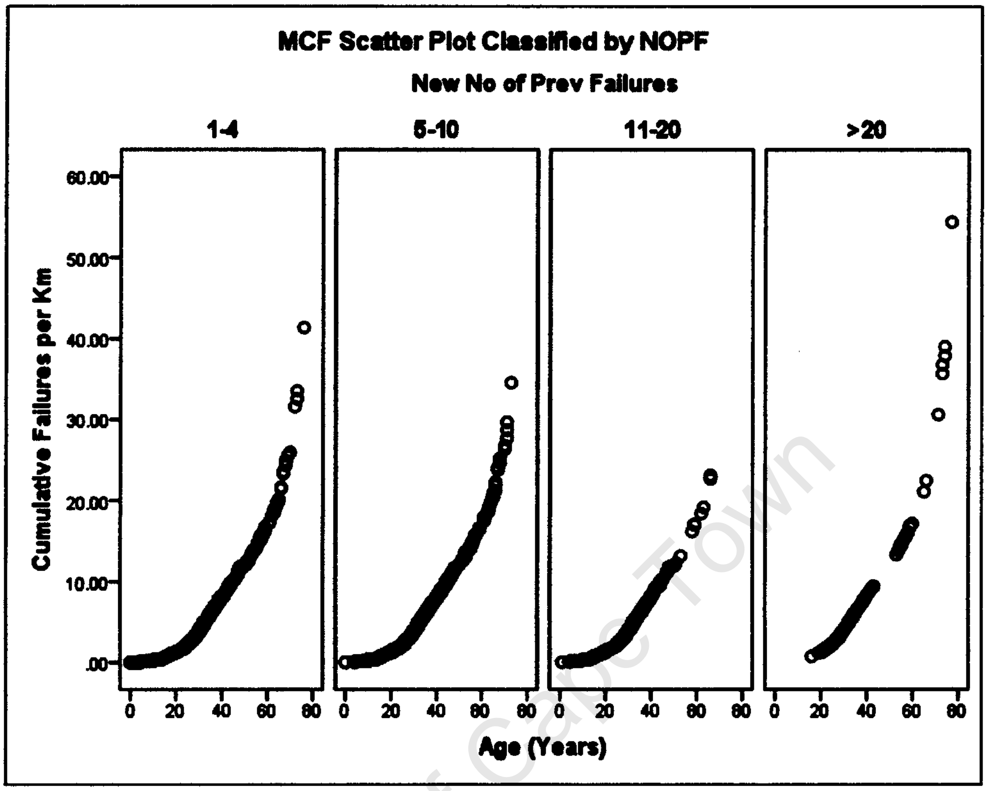


Figure 6-2: MCF estimates vs. age by NOPF

The distribution of the different categories of NOPF over time is illustrated below

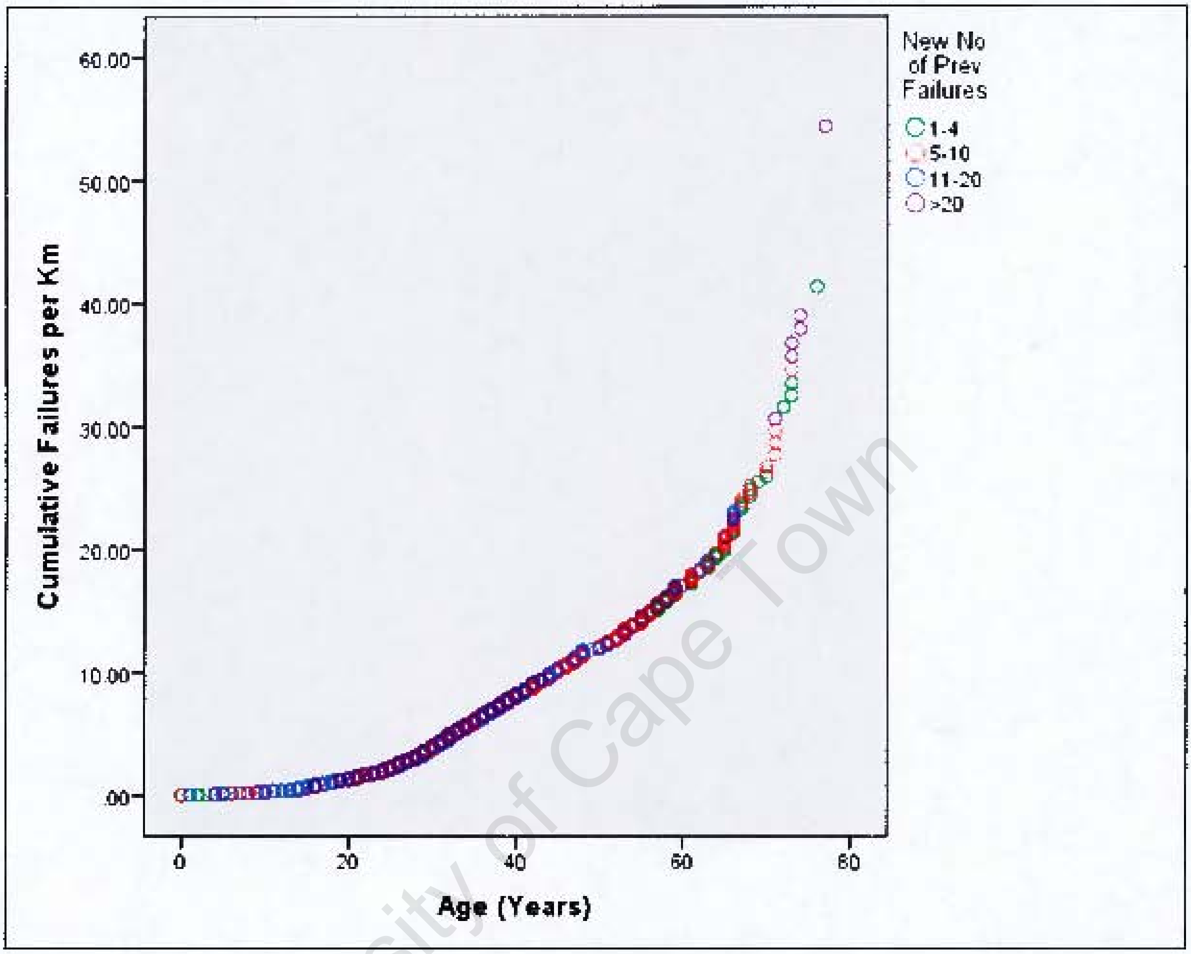


Figure 6-3: Distribution of different categories of NOFP over time

6.1.2 MCF Stratification by Installation Eras

The same procedure was conducted to capture any trends for the failure rate with different installation era. Plots of MCF estimate vs. survival times according to the predefined installation eras are shown in Figure 6-4.

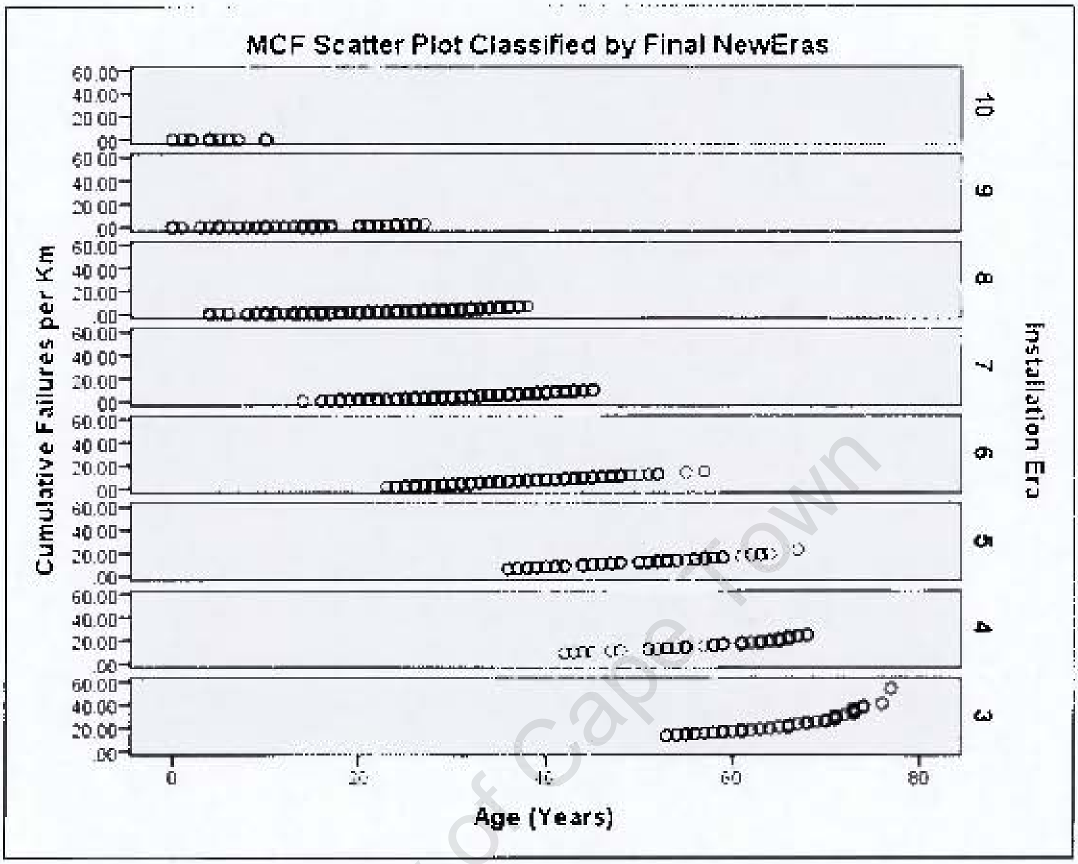


Figure 6-4: MCF estimates vs. age by installation eras1

No noticeable trends were depicted except for the first three eras. Installation eras have therefore been redefined into only three categories as shown in Table 6-1 and the modified plots are shown in Figure 6-5.

Table 6-1: Modified installation eras classification

Installation era	Modified Abbreviation
1930-1959	1
1960-1990	2
1990-2010	3

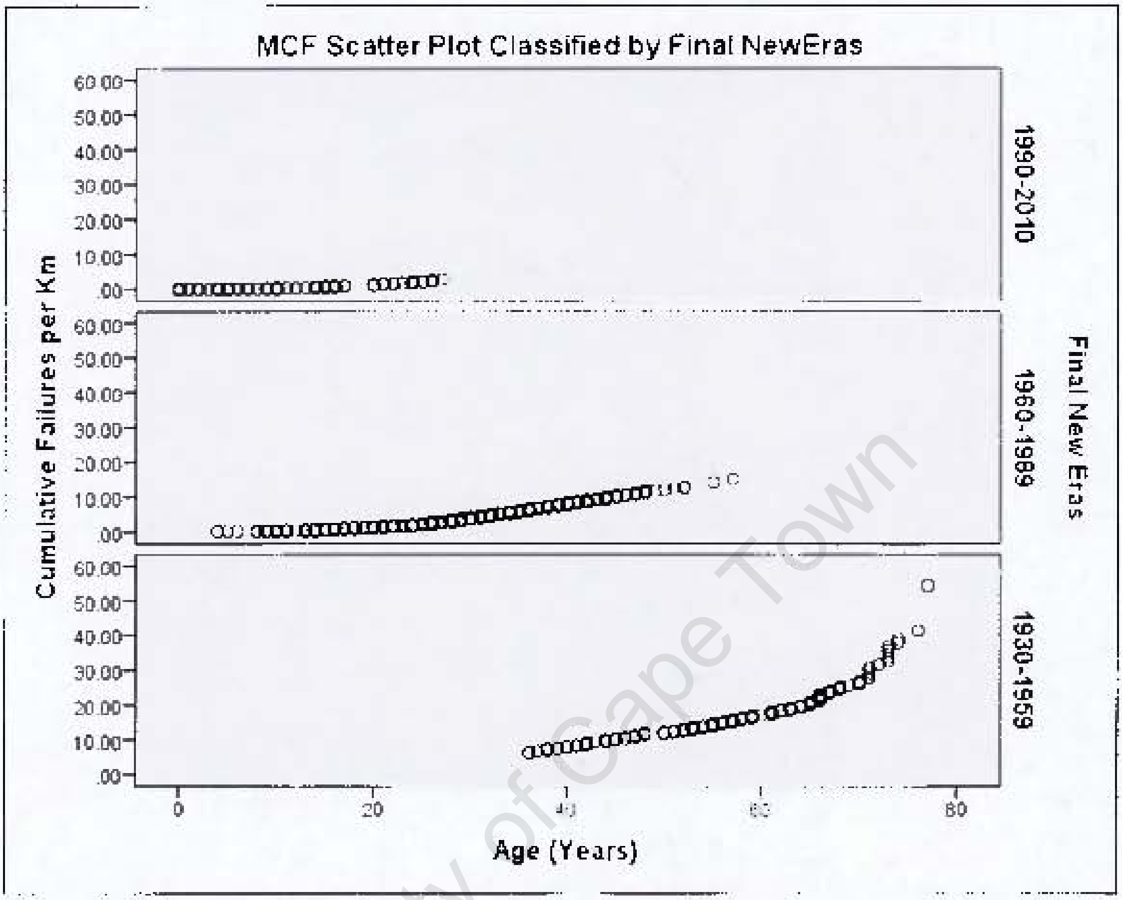


Figure 6-5: MCF estimates vs. age by final installation eras

It is clear that there is some noticeable difference in the FR (i.e. the slope of the MCF) for different Eras. Era 1 which denotes to pipes installed between 1930 and 1959, encountered the highest failure rates, then eras 2 and 3 respectively. This is quite reasonable as the older pipes are expected to be more prone to failure than the younger pipes. In terms of the bathtub curve, pipes in the first era are in the wear out stage, some of the pipes of the second era are still in the stabilized failure stage, and the rest seems to be at the beginning of the wear out stage while younger pipes of the third era are still in the in-service stage. The distribution of the different Eras over time is illustrated below

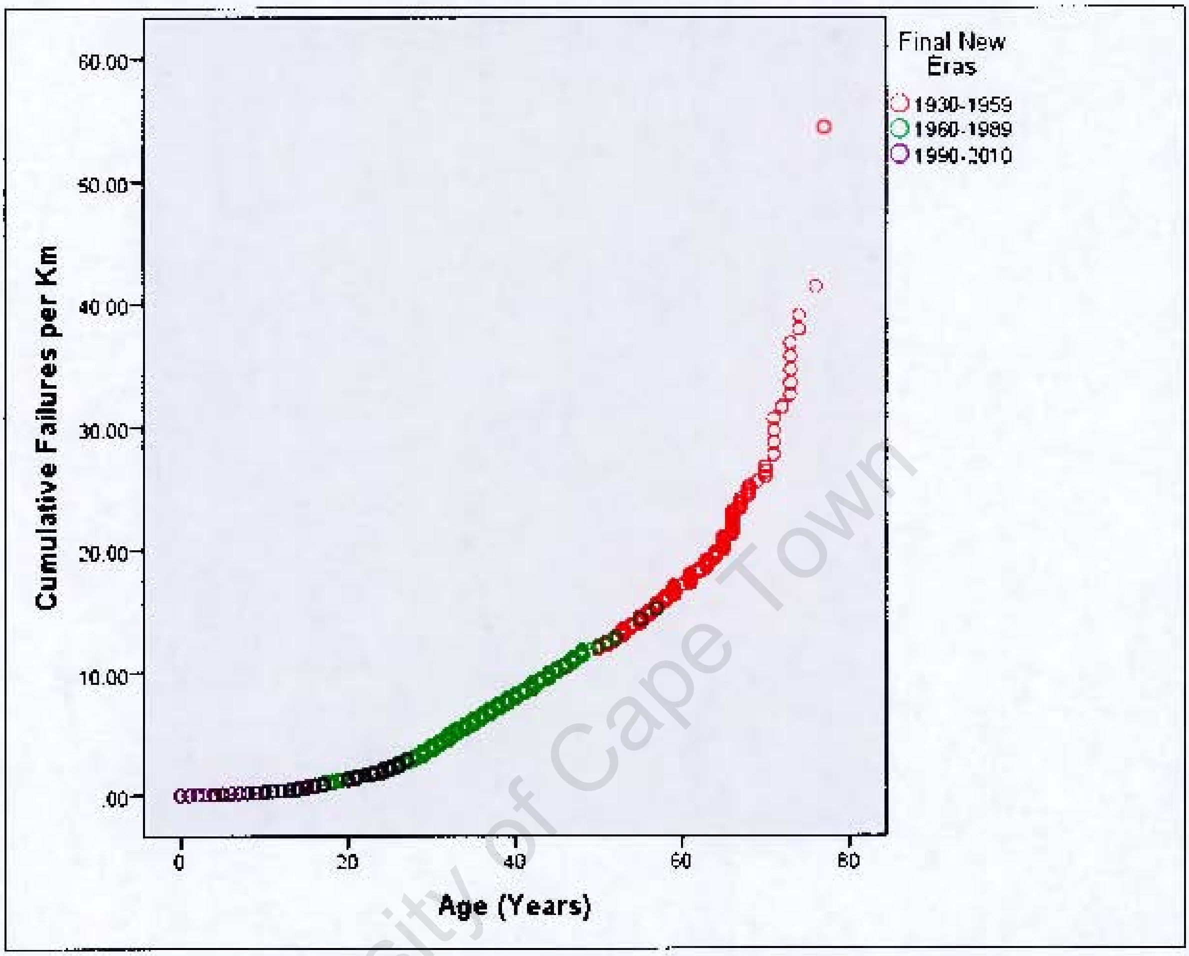


Figure 6-6: Distribution of different installation eras over time

6.2 The Failure Rate Models

In this section the MCF estimates was fitted to different types of curves and the best curve was derived using regression analysis. A different regression model was also derived for each distinct age stage defined in section 6.1 (i.e. multiple age regression models).

6.2.1 MCF Single Age Regression Analysis

The SPSS software has been used to fit the best curve for the failure rate. Different types of curves have been tested for goodness of fit and these are

- Linear model
- Quadratic model
- Cubic model
- Compound model
- Growth model
- Exponential

Results from fitting the above curves are illustrated as

Model Summary and Parameter Estimates

Dependent Variable: Cumulative Failures per Km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.869	8354.084	1	1254	.000	-7.003	.396		
Quadratic	.970	20200.087	2	1253	.000	.536	-.088	.007	
Cubic	.974	15708.227	3	1252	.000	-1.517	.159	-.001	7.116E-005
Compound	.865	8047.067	1	1254	.000	.268	1.083		
Growth	.865	8047.067	1	1254	.000	-1.318	.080		
Exponential	.865	8047.067	1	1254	.000	.268	.080		

The independent variable is Age (Years)

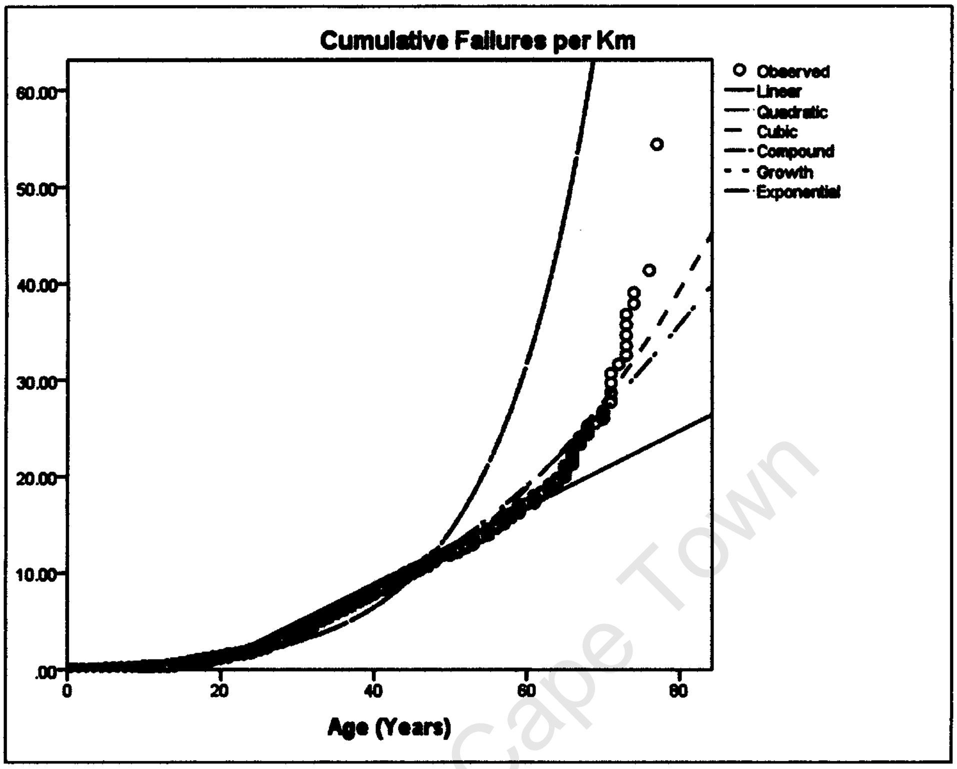


Figure 6-7: Curve fitting of the MCF as a function of age

The six models were found to be statistically significant and the adjusted R^2 for the six curves are statistically accepted. However, the quadratic model has been chosen to fit the due to its better F value. The model takes the form

$$\boxed{MCF\ (t) = b_2t^2 + b_1t + c}$$

And the final model for the MCF per km is given by

$$\boxed{MCF\ (t) = 0.007t^2 - 0.008t + 0.536}$$

The failure rate can then be derived by differentiating the above equation as

$$\boxed{\boldsymbol{FR(t) = \frac{d}{dt}MCF\,(t) = 0.014t - 0.008}}$$

While the regression analysis provides a good fit, Figure 6-7 suggests that there are three stages of failure in the life of pipes; and that a better fit could be obtained if these were analysed separately.

6.2.2 MCF Multiple Age Regression Analysis

Based on the results obtained from the MCF plots, three models have been proposed for each distinct stage of the bathtub curve shown in Figure 6-1, (i.e. multiple age stage regression). Again different models have been tested for goodness of fit for each age stage and results are shown as

6.2.2.1 Curve Fit for the First Age Stage (0 < Age < 25) years

The SPSS software has been used to develop the best regression model for ages <=25 years. The models that have been tested are

- Linear model
- Quadratic model
- Cubic model
- Compound model
- Growth model
- Exponential

Results from fitting the above models are

Model Summary and Parameter Estimates

Dependent Variable: Cumulative Failures per Km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.906	3220.459	1	336	.000	-.697	.104		
Quadratic	.993	23537.775	2	335	.000	.095	-.028	.004	
Cubic	.993	15714.085	3	334	.000	.079	-.022	.004	1.303E-005
Compound	.949	6285.102	1	336	.000	.050	1.173		
Growth	.949	6285.102	1	336	.000	-2.998	.160		
Exponential	.949	6285.102	1	336	.000	.050	.160		

The independent variable is Age (Years).

The six models were found to be highly significant, however it is apparent from the results displayed above that the quadratic and cubic model are the best models that fits the data. The quadratic model has been chosen as it produced better F value. It might, however, be more logical to force the constant in the equations to be zero, (see the graphs for the two models shown in Figure 6-8 and Figure 6-9).

Figure 6-8: The cubic regression model for ages <=25 years

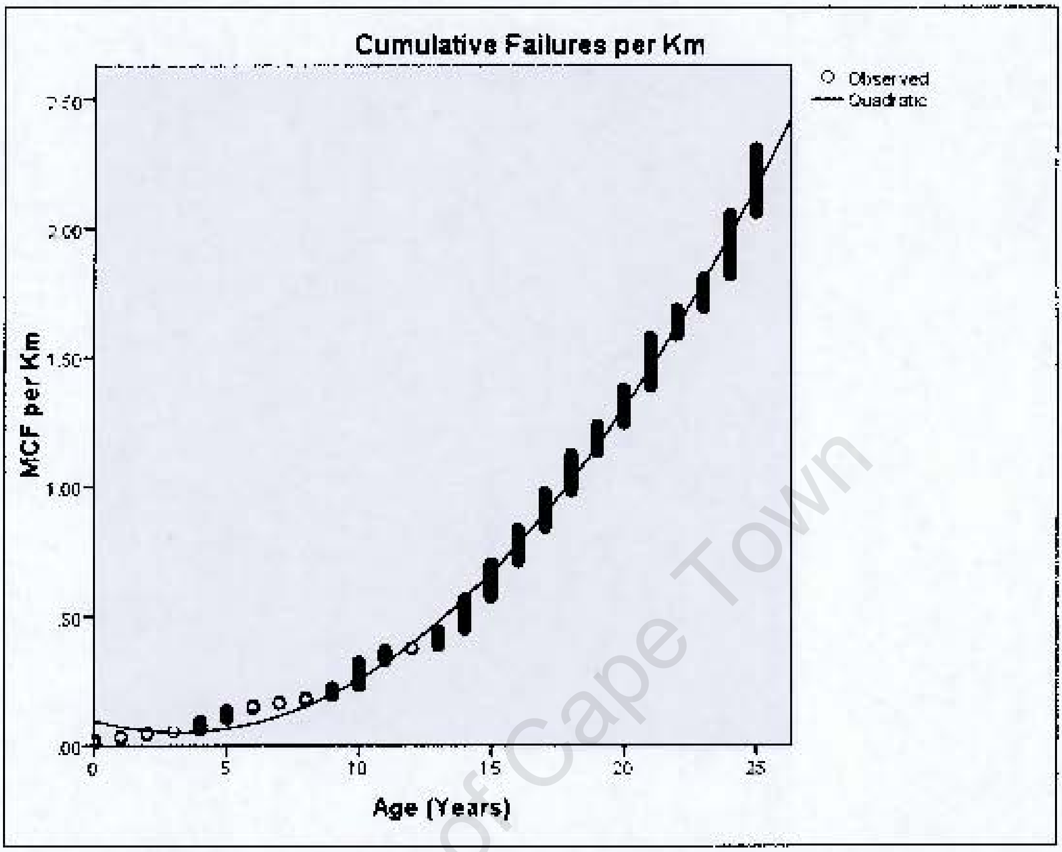


Figure 6-9: The quadratic regression model for ages <=25 years

6.2.3.1 Curve Fit for the Second Age Stage (25 < Age < 65 years)

The six models have been tested for goodness of fit for this age category. Results are demonstrated as

Model Summary and Parameter Estimates

Dependent Variable: Cumulative Failures per Km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.995	184423.940	1	880	.000	-9.025	.430		
Quadratic	.997	160692.366	2	879	.000	-5.968	.276	.0018021	
Cubic	.997	172190.416	2	879	.000	-6.883	.348	0E-7	1.441E-005
Compound	.935	12590.500	1	880	.000	.773	1.057		
Growth	.935	12590.500	1	880	.000	-.258	.055		
Exponential	.935	12590.500	1	880	.000	.773	.055		

The independent variable is Age (Years).

Results from fitting the curve shows that the linear, quadratic and cubic curves are the best curves that represent the data. Graphs for the three curves are illustrated in Figure 6-10, Figure 6-11and Figure 6-12.

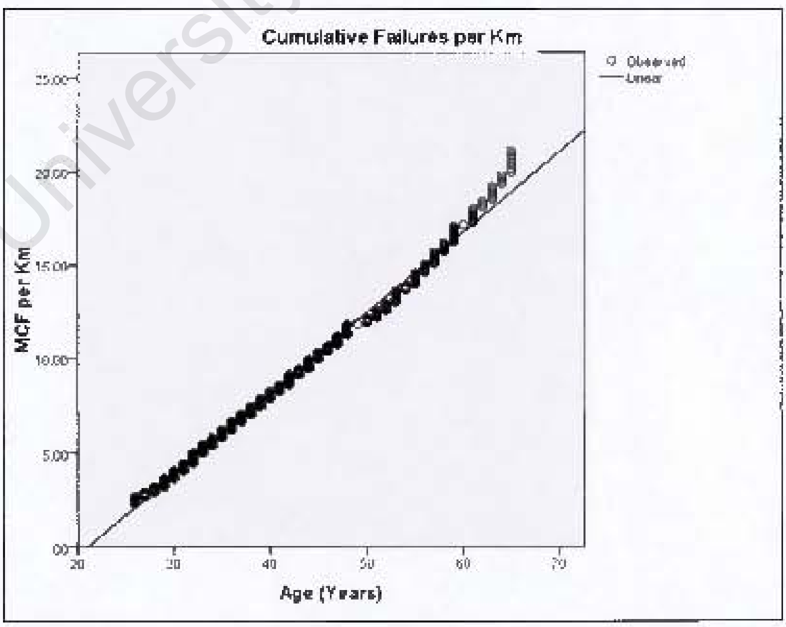


Figure 6-10: The linear regression model for 25< age < 65 years

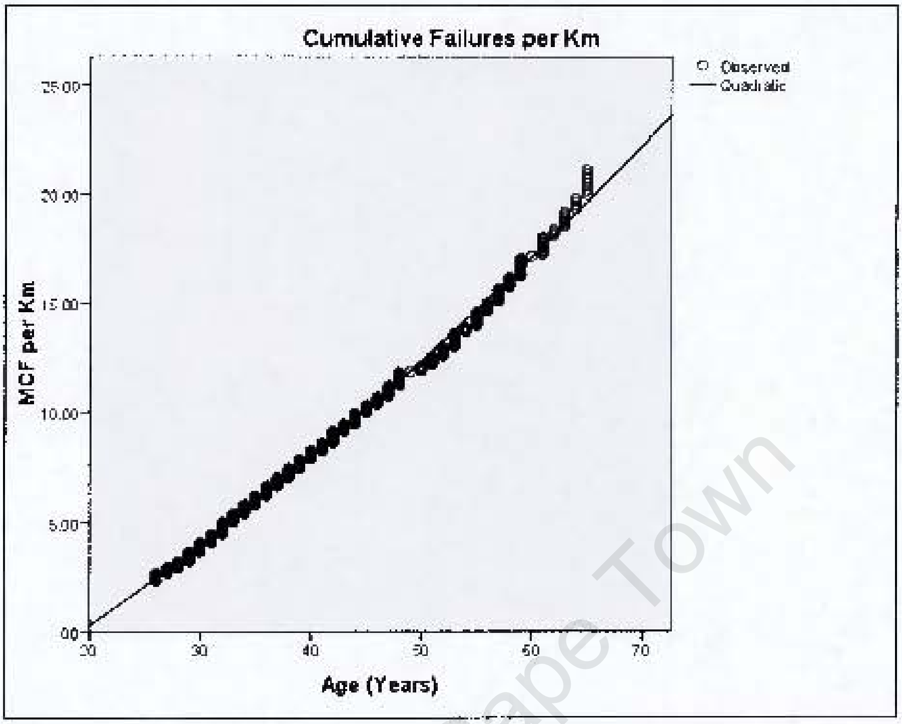


Figure 6-11: The quadratic model for 25< age < 65 years

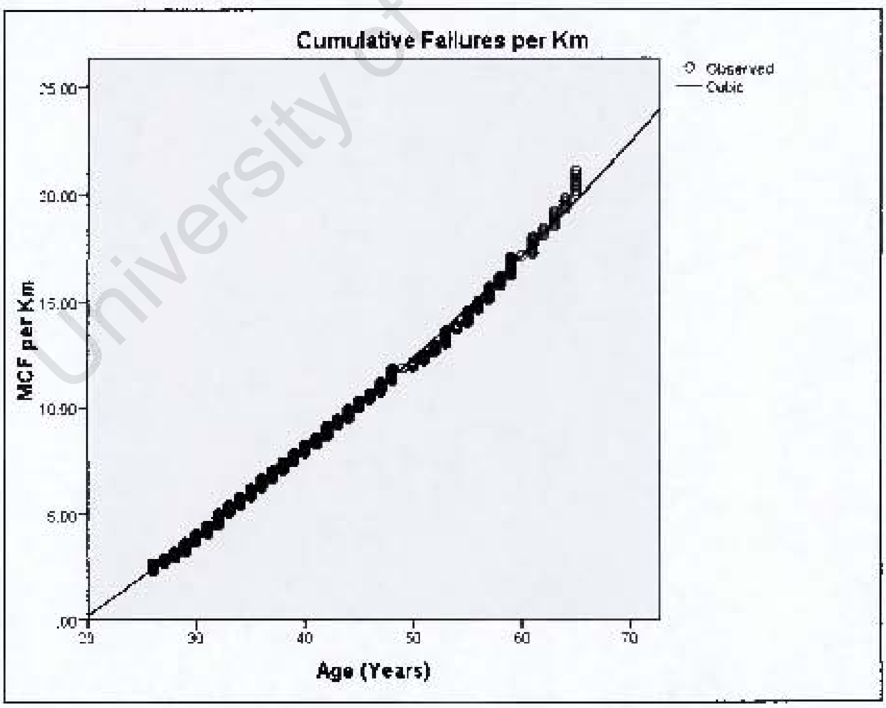


Figure 6-12: The cubic model for 25 < age < 65 years

The linear model has been chosen to fit the curve as it has a better F value although its R Square is slightly worse. Figure 6-10 reveals linear relationship between the MCF estimates and the age of pipes until age 60 years where the curve starts to depart from linearity. This suggests that the data for pipes older than 60 years should be allocated to the third age stage. The model has then been fitted for the period 25 < age < 60 years and results are demonstrated as follows

6.2.3.2 Curve Fit for the Second Age Stage (25 < Age < 60 years)

Model Summary and Parameter Estimates

Dependent Variable: Cumulative Failures per Km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.997	322746.593	1	851	.000	-8.611	.417		
Quadratic	.998	185347.104	2	850	.000	-7.330	.351	.001	
Cubic	.998	186707.733	2	850	.000	-7.731	.383	.000	6.767E-006
Compound	.937	12699.341	1	851	.000	.685	1.060		
Growth	.937	12699.341	1	851	.000	.379	.059		
Exponential	.937	12699.341	1	851	.000	.685	.059		

The independent variable is Age (Years).

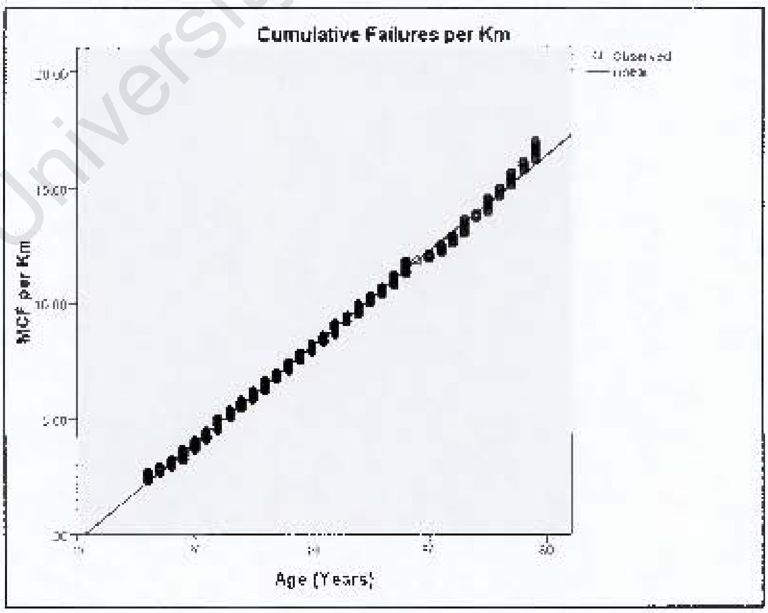


Figure 6-13: The linear model for 25 < Age < 60 years

Curve Fit for the third Age Stage (Age > 60 years)

The six curves have been again tested for goodness of fit for this age stage and the results are illustrated as

Model Summary and Parameter Estimates

Dependent Variable: Cumulative Failures per Km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.883	473.475	1	63	.000	-81.016	1.579		
Quadratic	.968	946.165	2	62	.000	415.463	-13.157	.109	
Cubic	.971	1022.805	2	62	.000	121.609	.000	-.087	.001
Compound	.961	1540.647	1	63	.000	.461	1.061		
Growth	.961	1540.647	1	63	.000	-.775	.059		
Exponential	.961	1540.647	1	63	.000	.461	.059		

The independent variable is Age (Years).

Clearly, that the quadratic and the cubic model are the best for fitting the data however the cubic has been chosen as it produced a better adjusted R square and F value. Graphs for the two models are illustrated as

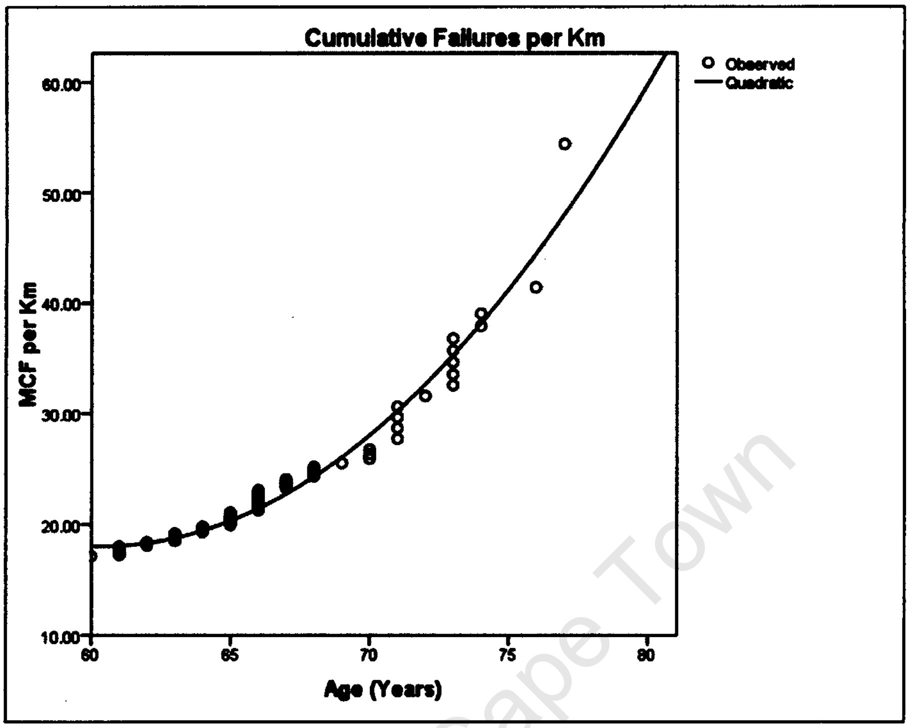


Figure 6-14: The quadratic model for Ages > 60

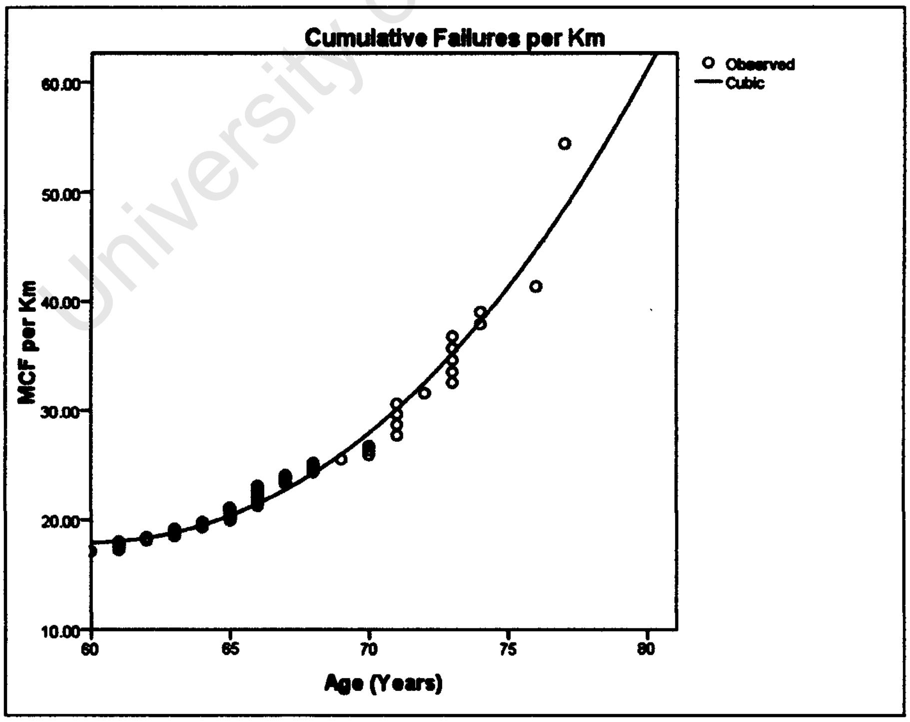


Figure 6-15: The cubic model for Age > 60 years

Table 6-2: Final MCF regression models summary

Final MFC Regression Models Summary

Model Name	Quadratic for 0 < Age <25 years		Linear for 25 < Age <60 years		Cubic Model for Age > 60 years	
Dependent Variable	Cumulative Failures/Km		Cumulative Failures/Km		Cumulative Failures/Km	
Total Cases	338		853		65	
Adjusted R Square	.993		.997		.970	
Independent variable	B	t	B	t	B	t
b_3	-	-	-	-	.001	12.194
b_2	.004	64.375	-	-	-.087	-10.733
b_1	-.028	-13.401	.417	568.108	-	-
(Constant)	0	6.196	-8.611	-311.317	121.609	9.845

Final MFC Regression Models Summary

- The Quadratic Model for Age < = 25 years:

$$\boldsymbol{MCF_1(t) = 0.004t^2 - 0.028t}$$

- The Linear Model for Age 25 < Age < 60 years:

$$\boldsymbol{MCF_2(t) = 0.417t - 8.611}$$

- The Cubic Model for Age > 60 years:

$$\boldsymbol{MCF_3(t) = 0.001t^3 - 0.087t^2 + 121.609}$$

The Failure Rate:

- For Age < = 25 years:

$$\boldsymbol{FR_1(t) = \frac{d}{dt}MCF_1(t) = 0.008t - 0.028}$$

- For Age 25 < Age < 60 years:

$$\boldsymbol{FR_2(t) = \frac{d}{dt}MCF_2(t) = 0.417}$$

- For Age > 60 years:

$$\boldsymbol{FR_3(t) = \frac{d}{dt}MCF_3(t) = 0.003t^2 - 0.174t}$$

6.3 Application of the PHM

The proposed methodology in section 5.3.2.5 was applied to the case study dataset. Analysis of the PHM has been conducted using the variables primarily selected and shown in Table 5-14.

6.3.1 Descriptive Statistics

Descriptive statistics has been derived to describe the distribution of the primary selected covariates in general and within the identified age strata from the MCF analysis. It is useful to note that the categories used for this descriptive statistics do not necessarily represent the categories used throughout the analysis of the PHM. Different categorization might be undertaken according to specific statistical requirements or results. Results are illustrated as follows

Frequency tables

The following tables describe the previously defined explanatory variables in terms of frequency and percentage.

Agestrata2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 - 25	471	26.7	26.7	26.7
	25 - 60	1180	67.0	67.0	93.7
	> 60	111	6.3	6.3	100.0
	Total	1762	100.0	100.0	

Length Categories: For the purpose of the descriptive statistics, the length of pipes has been classified into the following four categories. Other classification will be made for pipe lengths for the purpose of the application of the PHM.

LCats		Frequency	Percent	Valid Percent	Cumulative Percent
1	1 - 200 m	500	28.4	28.4	28.4
2	201 - 400 m	565	32.1	32.1	60.4
3	401 - 1000 m	587	33.3	33.3	93.8
4	> 1000 m	110	6.2	6.2	100.0
	Total	1762	100.0	100.0	

Censorship

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 (censored)	506	28.7	28.7	28.7
	1 (failed)	1256	71.3	71.3	100.0
	Total	1762	100.0	100.0	

New NOPF

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 (1-4)	731	41.5	41.5	41.5
	2 (5-10)	648	36.8	36.8	78.3
	3 (11-20)	260	14.8	14.8	93.0
	4 (> 20)	123	7.0	7.0	100.0
	Total	1762	100.0	100.0	

Final New Eras

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 (1930-1959)	270	15.3	15.3	15.3
	2 (1960-1980)	1355	76.9	76.9	92.2
	3 (> 1980)	137	7.8	7.8	100.0
	Total	1762	100.0	100.0	

Graphs

NOPF by Age Strata

This plot describes the distribution of the number of previous failures within each age stratum. The plot shows that the second age category represents the vast majority of pipes in the network. Most pipes in the three age categories incurred from 1 to 4 failures. The next in turn is the second and third categories of the NOPF (i.e. 5-10 and 11-20 failures). Exception of this pattern is the last age category which incurred higher number of pipes with more than 20 failures than the other two age categories.

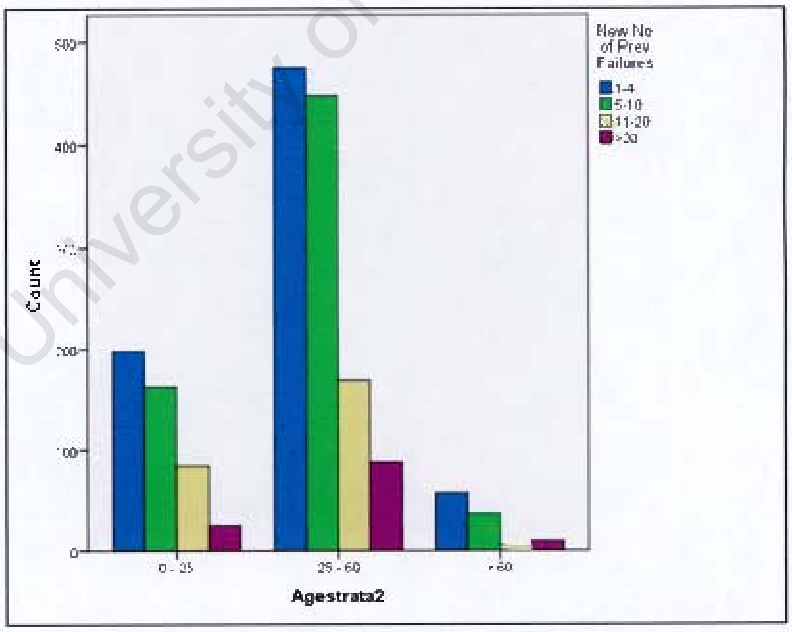


Figure 6-16: Distribution of the NOPF within the three age strata.

Length of pipes by Age Strata

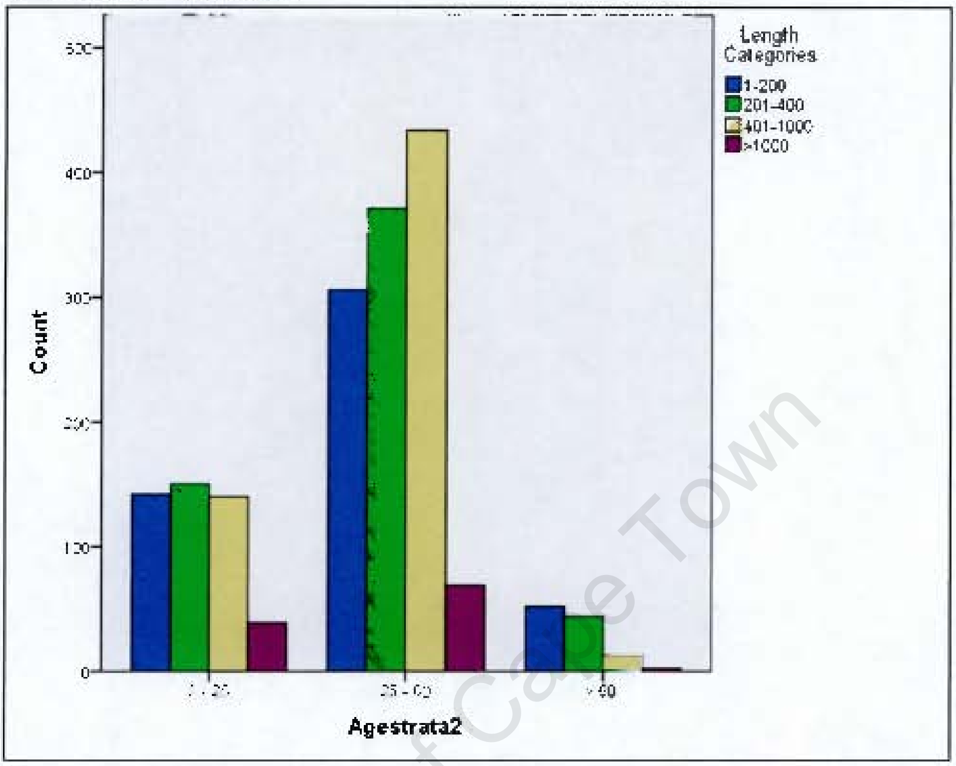


Figure 6-17: Distribution of length within the three age strata

Figure 6-17 tells that pipes of the first age strata have similar distribution of pipe length categories except for long pipes which only represents around 40 pipes out of the 471 pipes in this category. The second and third categories of pipes have a slightly different distribution of pipe's length. Most of pipes in the second age category have pipe length between 400 and 1000m, whereas most of the pipes of the third age category have length that range between 1-200m.

Installation eras of pipes by Age Strata

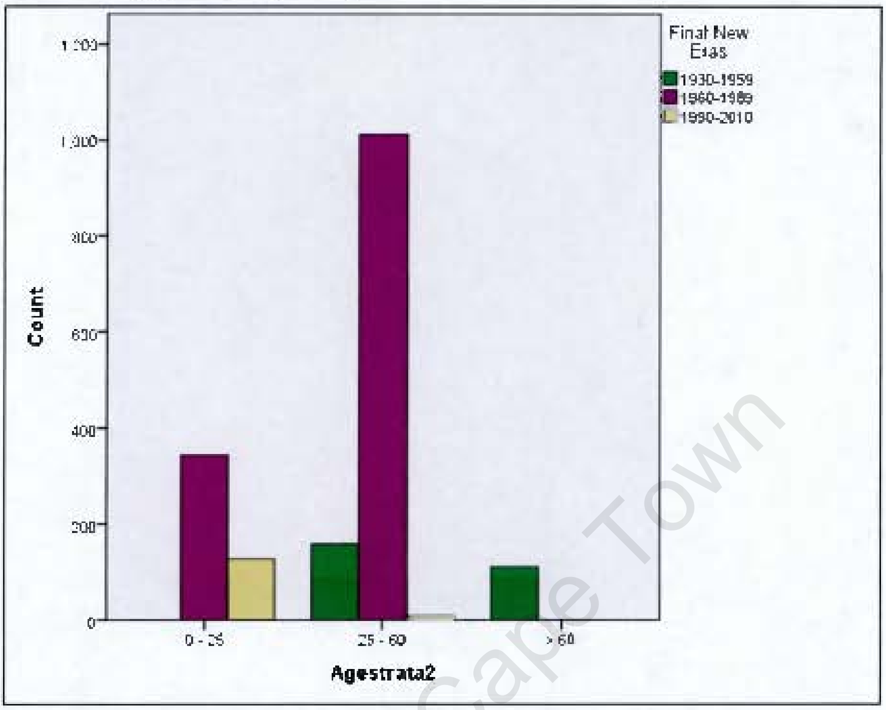


Figure 6-18: Distribution of installation eras by age strata

Figure 6-18 shows that more than 300 pipes of the first age category were installed between 1960 and 1980 and the rest of pipes of this category (i.e. around 150 pipes) were installed after 1980 with no pipes of this category were installed before 1960. The vast majority of pipes in the second age category were installed between 1960 and 1980 with less than 200 pipes out of the 1226 pipes in this age category were installed before 1960 and a minor number of pipes were installed after 1980. Pipes in the third age category were all installed before 1960.

Censorship of pipes by Age Strata

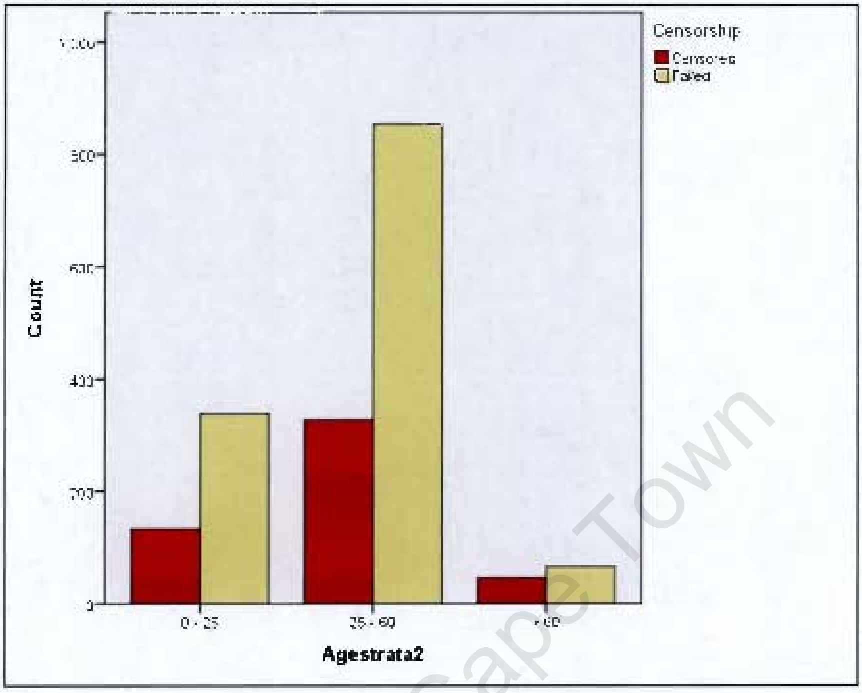


Figure 6-19: Distribution of censorship by age strata

Figure 6-19 shows that the number of failing pipes in the first age category is more than double the number of censored pipes, and so does the case for pipes in the second age category whereas the number of censored pipes is almost equal to the number of failed pipes in the third age category.

The procedure of variables selection for the final PHM for the case study dataset will be applied in the next sections in order to accurately identify the effect of specific covariates on the hazard of failure of pipes and the various stages of the building of the final model will be presented.

6.3.2 Application of the COX's PHM at Pipe Level

Several investigatory PHMs have then been applied to the whole dataset to examine the statistical significance of the primary selected variables. Each variable has been individually checked for significance and a PHM was developed for each variable that found to be statistically significant given that it satisfies the proportional hazard assumption. In adoption of the step wise selection method of a set of explanatory variables, variables will be added to the model once at a time and the statistical significance of them will be examined at each step as well as the proportionality assumption. The best combination of a set of explanatory variables was used to develop a PHM to assess the multiplicative effect of covariates on the hazard of pipes. The included covariates at this stage are

1. The total length of pipes (L)
2. The total number of previous failures (NOPF)
3. The final defined installation eras (NewEras2, dummy variable)

The SPSS software have been used to carry out the analysis and the output results are illustrated as

6.3.2.1 The PHM by Including the Variable Length (L)

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
16474.583	53.694	1	.000	47.820	1	.000	47.820	1	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.001	.000	53.485	1	.000	1.001	1.000	1.001

Results from running the PHM to the variable length revealed that the variable is highly statistically significant. The proportionality assumption was tested for this variable using the Schoenfeld Residuals method which is described in details in (Kleinbaum & Klein, 2005).

The correlations result for the partial residual for L was found to be significant which implies that the proportionality assumption for this variable is violated for this model.

Correlations of the partial residual of L

		Partial residual for L	Rank of age
Partial residual for L	Pearson Correlation	1	.111**
	Sig. (2-tailed)		.000
	N	1256	1256
Rank of age	Pearson Correlation	.111**	1
	Sig. (2-tailed)	.000	
	N	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

The analysis of the PHM for this variable should, therefore, consider the time effect of the variable L and apply the Extended Cox Model. Further investigation for the applicability of the PHM to the dataset has been conducted and by examining the statistical significance for each of the other two explanatory variables as well as combination of them (i.e. NOPF and Final New Fras2)

6.3.2.2 The PHM by Including the Variable NOPF

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
16433.938	101.572	1	.000	88.465	1	.000	88.465	1	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
NOPF	.037	.004	99.450	1	.000	1.038	1.030	1.045

Results from running the model shows that the variable NOPF is highly significant. The proportionality assumption has been examined using the same procedure of the Schoenfeld Residuals for this variable and results shows that the partial residual for the variable NOPF is insignificant which implies that the proportionality assumption is satisfied. The correlation table for the partial residual for the variable NOPF is shown below

Correlations of the partial residual of NOPF

		Partial residual for NOPF	Rank of age
Partial residual for NOPF	Pearson Correlation	1	.001
	Sig. (2-tailed)		.960
	N	1256	1256
Rank of age	Pearson Correlation	.001	1
	Sig. (2-tailed)	.960	
	N	1256	1256

Graphs for survival, one minus survival, log minus log, and hazard functions at mean of the variable NOPF are illustrated as follows

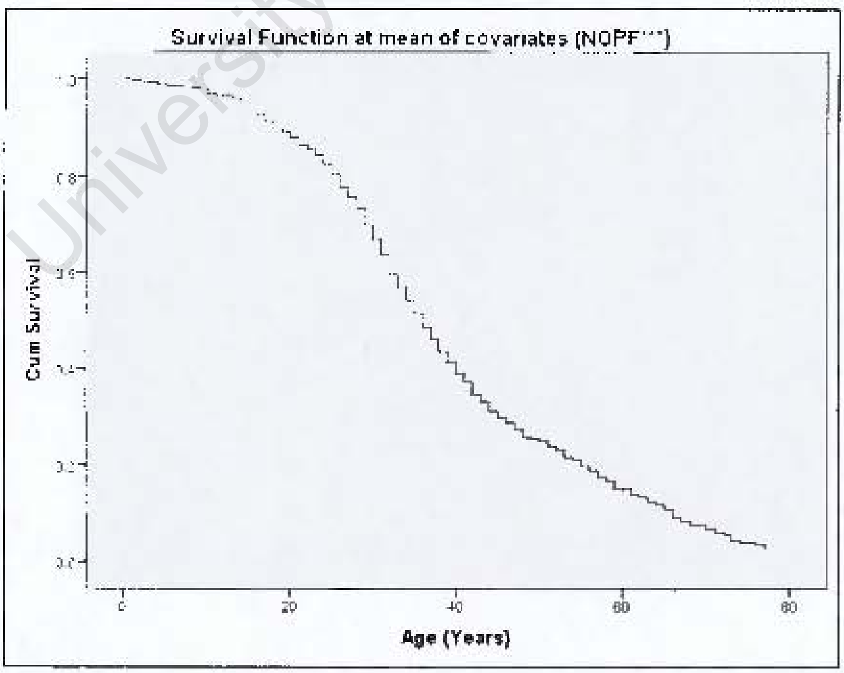


Figure 6-20: Survival function at mean of covariates (PHM, NOPF)

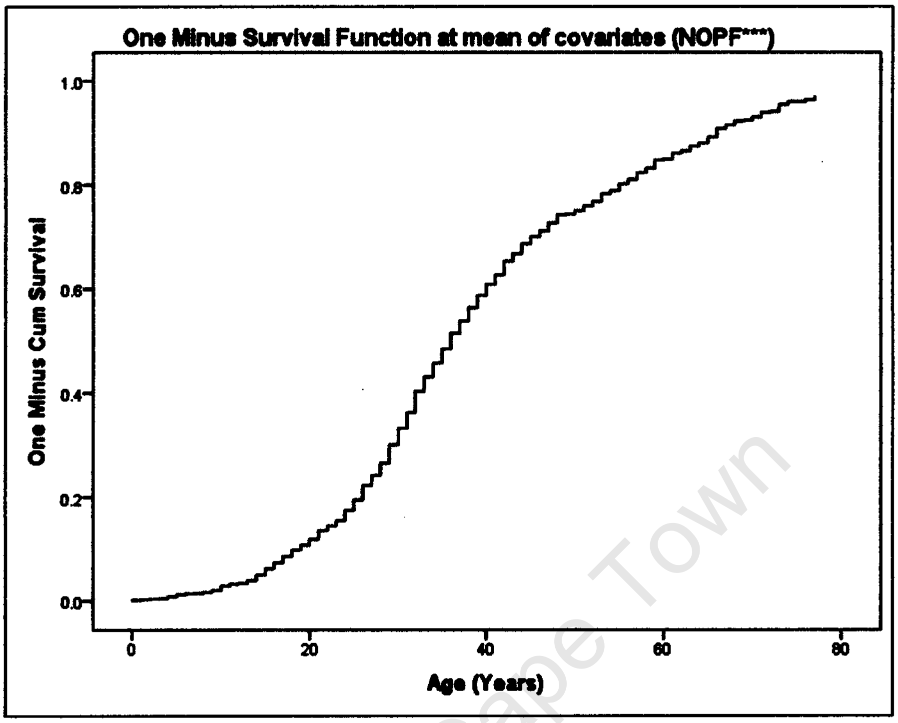


Figure 6-21: One minus survival function at mean of covariates (PHM, NOPF)

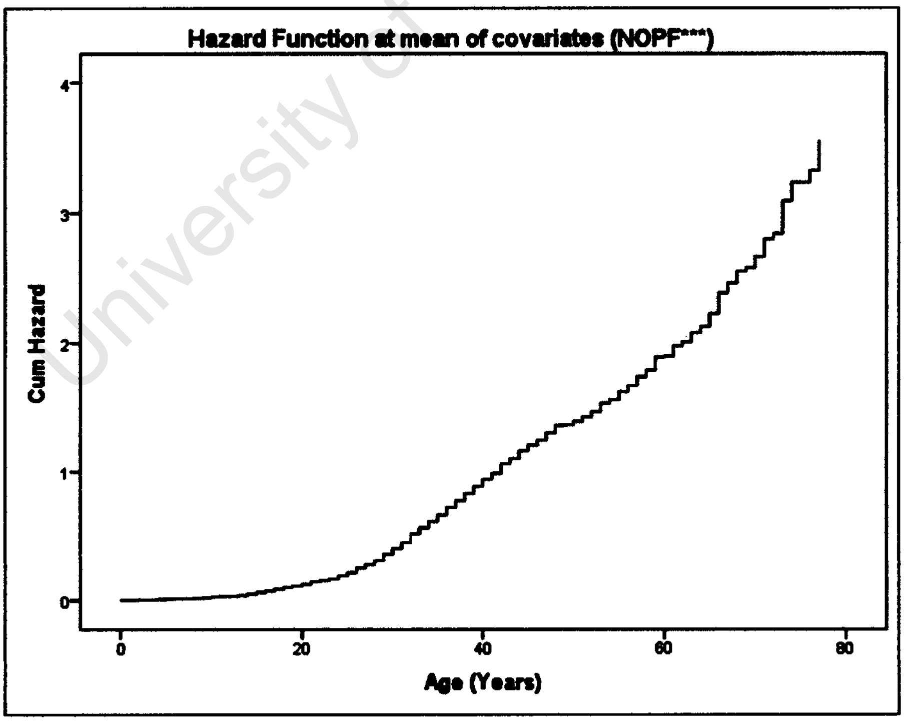


Figure 6-22: Hazard function at mean of covariates (PHM, NOPF)

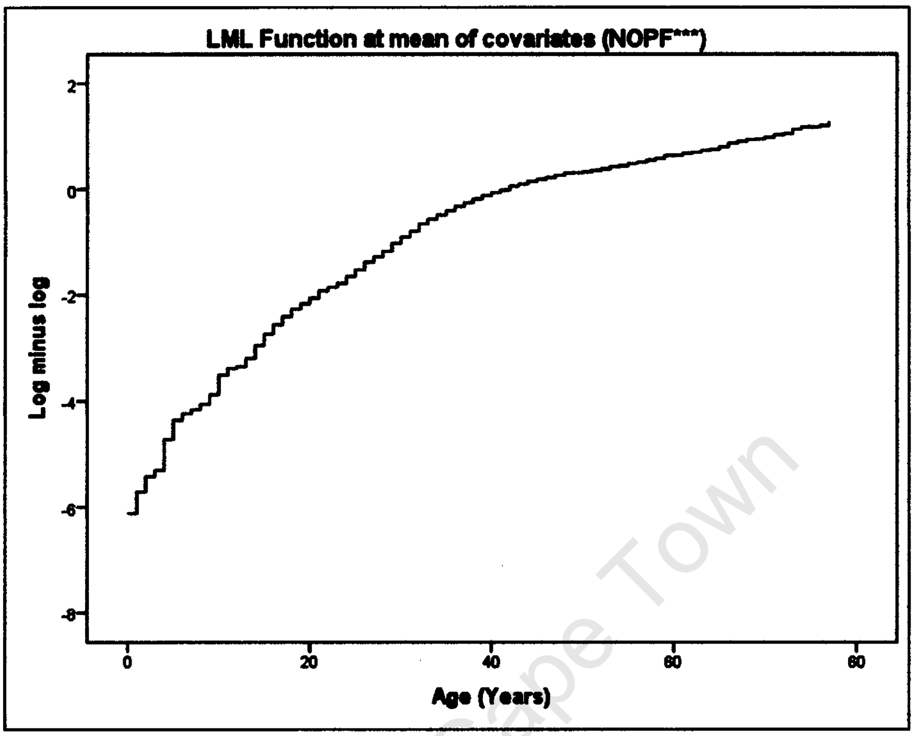


Figure 6-23: Log minus Log function at mean of covariates (PHM, NOPF)

From the above figures one can notice that the probability of survival for individuals drops for ages between 25 – 45 years during which the hazard function also increases. Pipes then start the wear out stage where the hazard function increases rapidly with many ties indicating higher occurrences of failure from that age on.

Estimation of the baseline hazard function (PHM, NOPF)

The baseline hazard function has then been estimated using regression analysis. The estimated baseline cumulative hazards obtained by the SPSS software, shown in the appendix, were plotted against the survival age of pipes. The scatter plot of these two variables is shown below

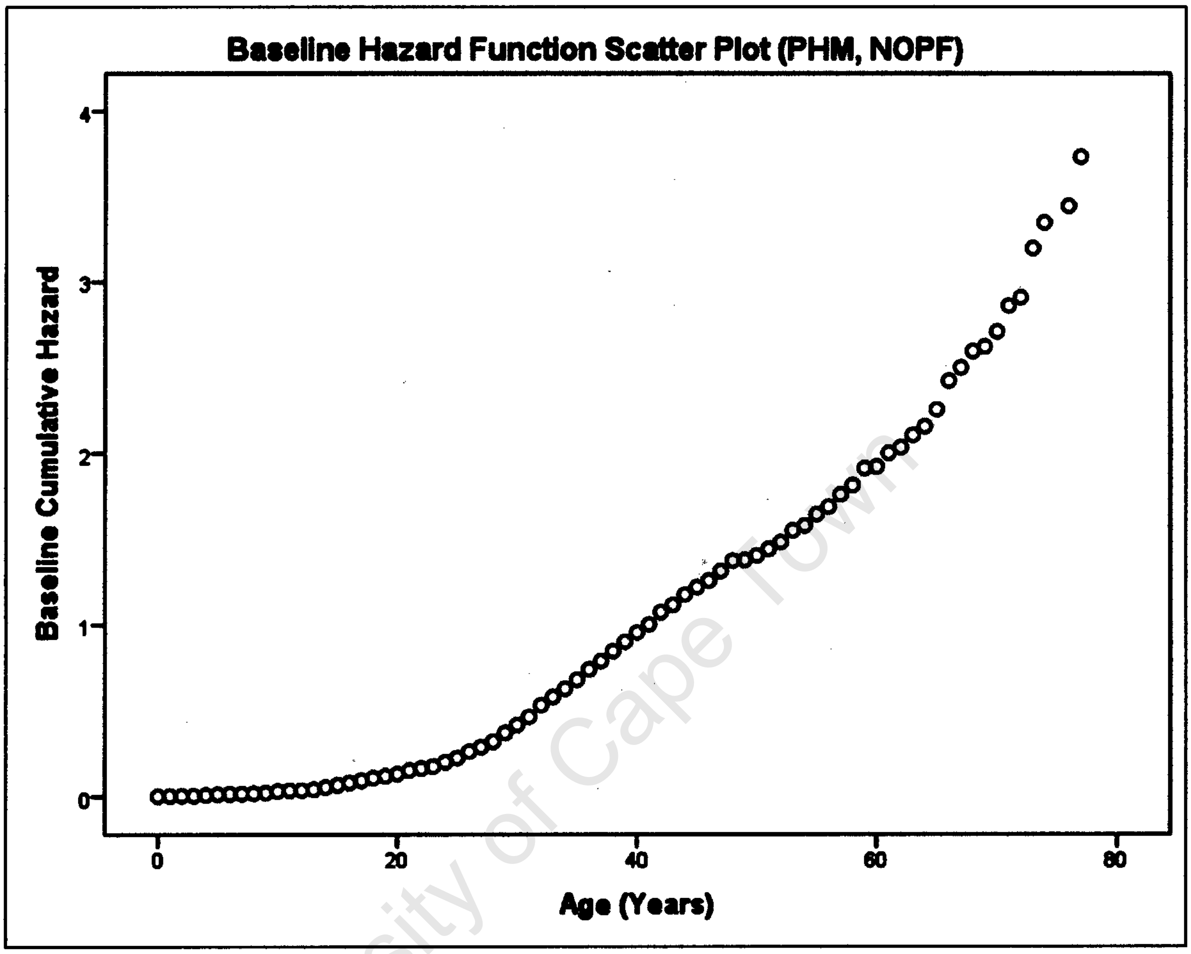


Figure 6-24: Baseline hazard function scatter plot (PHM, NOPF)

One can easily distinguish between three distinct stages for the baseline hazard over time. The first stage for ages between 0 – 30 years where the baseline hazard increases stably followed by a steep increase in the hazard for ages between 30 and 50 years which represent the second stage and then continue with much higher hazard rates for ages greater than 50 years. The curve can be fitted for the whole age span as well as individually for each stage for more accuracy. Different types of curves have been examined for goodness of fit for the different age stages and results are illustrated as

Curve fit for BCH function (0 < Age < 30 years) (PHM, NOPF)

Model Summary and Parameter Estimates

Dependent Variable: Baseline Cumulative Hazard

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.848	162.082	1	29	.000	-.067	.012		
Quadratic	.989	1209.199	2	28	.000	.021	-.006	.0061	
Cubic	.997	3076.193	3	27	.000	-.002	.004	.0003	1.922E-005
Compound	.966	822.708	1	29	.000	.004	1.176		
Growth	.966	822.708	1	29	.000	-5.413	.163		
Exponential	.966	822.708	1	29	.000	.004	.163		

The independent variable is Age (Years).

Both the quadratic and cubic models were found to better fit the data; however the cubic model produced a better R square and F values so this model has been chosen to fit the data for this range of age.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age (Years)	.004	.001	.299	3.079	.005
Age (Years) ** 2	.0003	.000	-.609	-2.581	.016
Age (Years) ** 3	1.922E-005	.000	1.325	8.884	.000
(Constant)	-.002	.004		-.572	.572

The graph of the model that fits the curve is illustrated as

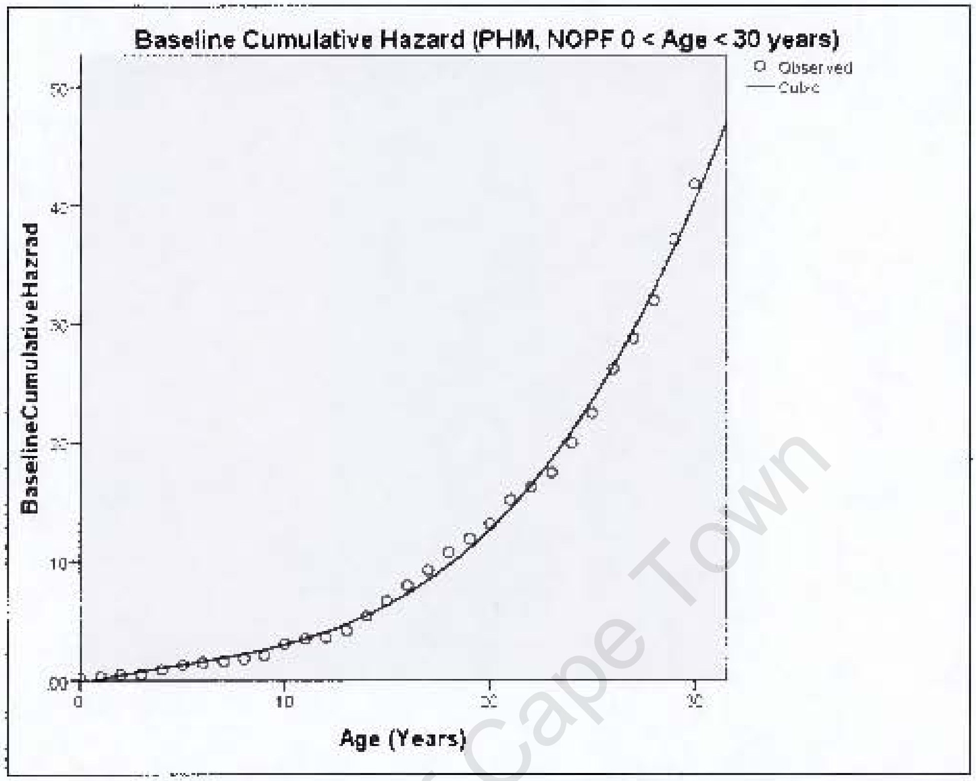


Figure 6-25: Baseline hazard function (PHM, NOPF, 0 < Age < 30)

Curve fit for BCH function (30 < Age < 50 years) (PHM, NOPF)

Model Summary and Parameter Estimates

Dependent Variable: Baseline Cumulative Hazard

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig	Constant	b1	b2	b3
Linear	.996	4522.104	1	18	.000	-1.110	.051		
Quadratic	.996	4505.292	2	17	.000	-1.845	.088	.0005	
Cubic	.998	4816.791	2	17	.000	-1.618	.071	.0000	-3.871E-006
Compound	.966	508.597	1	18	.000	.092	1.059		
Growth	.966	508.597	1	18	.000	-2.385	.057		
Exponential	.966	508.597	1	18	.000	.092	.057		

The independent variable is Age (Years).

Results from fitting the six above curves shows that the cubic model is the best that fits the data.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age (Years)	.071	.004	1.372	16.779	.000
Age (Years) ** 3	-3.871E-006	.000	-.377	-4.611	.000
(Constant)	-1.618	.112		-14.415	.000

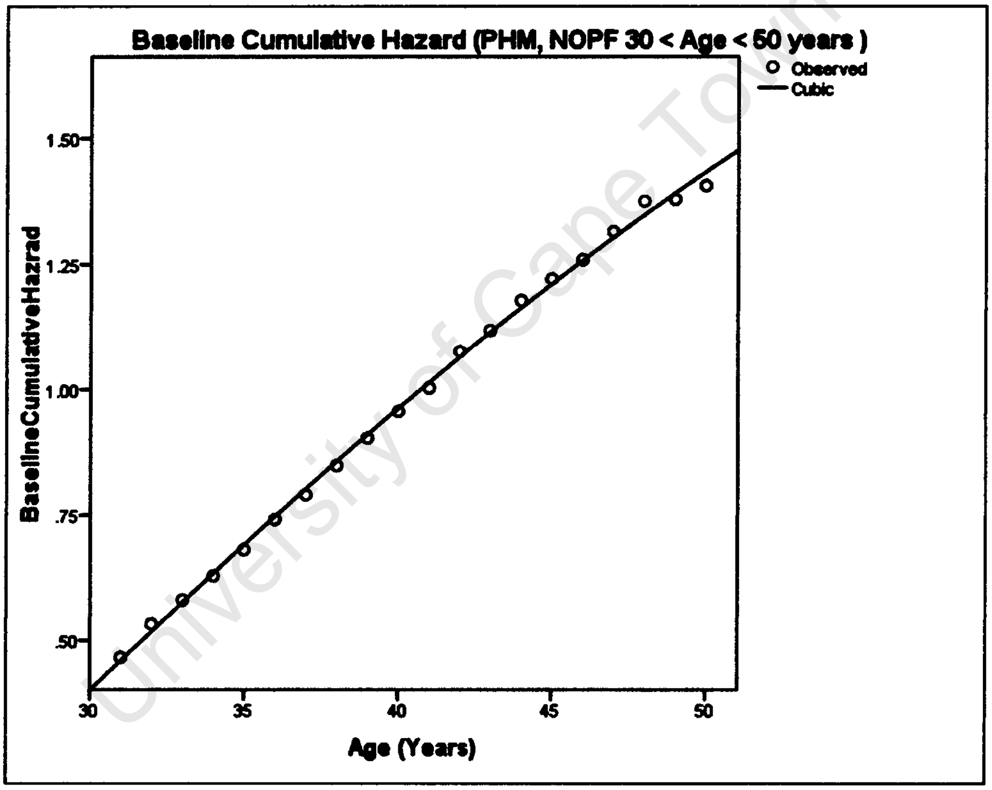


Figure 6-26: Baseline hazard function (PHM, NOPF, 30 < Age < 50)

Curve fit for BCH function (Age > 50 years) (PHM, NOPF)

Model Summary and Parameter Estimates

Dependent Variable: Baseline Cumulative Hazard

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.961	593.937	1	24	.000	-2.944	.082		
Quadratic	.994	1943.480	2	23	.000	5.741	-.194	.002	
Cubic	.995	2147.988	2	23	.000	1.710	.000	-.001	1.617E-005
Compound	.993	3570.636	1	24	.000	.230	1.036		
Growth	.993	3570.636	1	24	.000	-1.470	.036		
Exponential	.993	3570.636	1	24	.000	.230	.036		

The independent variable is Age (Years).

It is apparent that the cubic model is the best for fitting the data for this age stage.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age (Years) ** 2	-.001	.000	-1.388	-4.958	.000
Age (Years) ** 3	1.617E-005	.000	2.381	8.501	.000
(Constant)	1.710	.249		6.882	.000

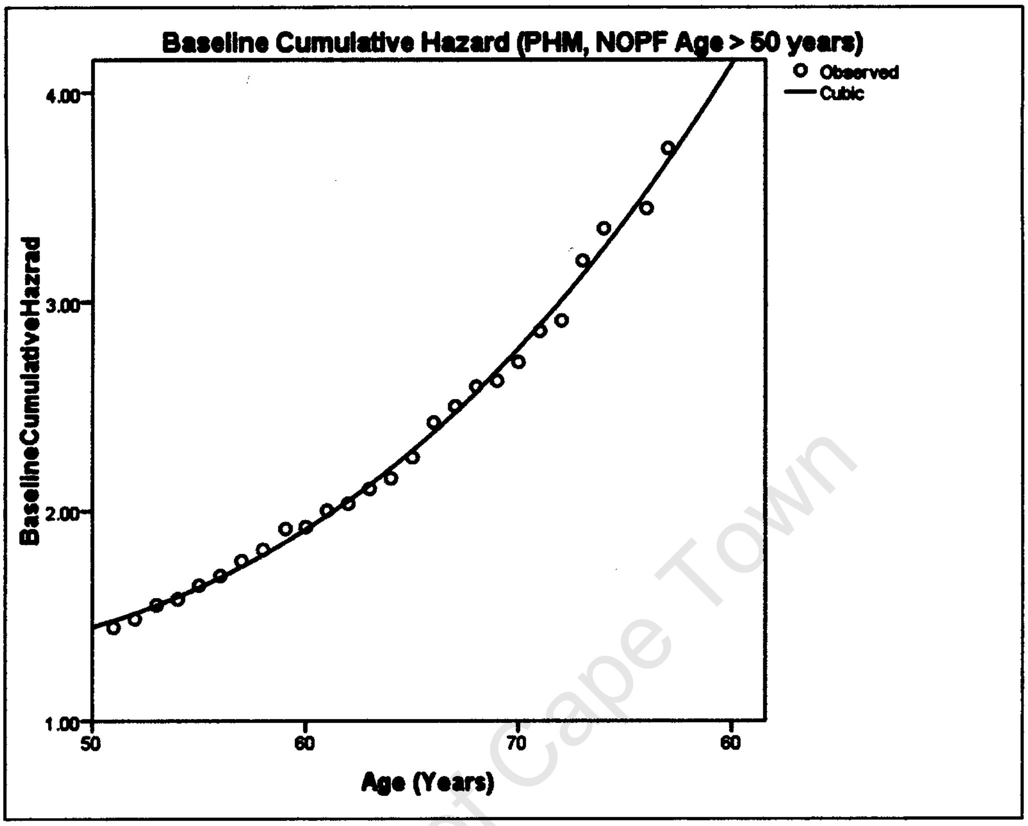


Figure 6-27: Baseline hazard function (PHM, NOPF; Age > 50)

The final PHM with NOPF is the only covariate is illustrated in Table 6-3.

Table 6-3: PHM at pipe level with NOPF as a covariate

Model Terms	**Mathematical expression**
The time independent term	$\mathbf{exp(0.037 \times NOPF)}$
Baseline hazard function (0 < Age < 30) years: Cubic	$\mathbf{h_{01} = (1.922E - 005)t^3 - 0.0003t^2 + .0.004t - 0.002}$
Baseline hazard function (30 < Age < 50) years: Cubic	$\mathbf{h_{02} = -(3.871E - 006)t^3 + 0.071t - 1.618}$
Baseline hazard function (Age > 50) years: Cubic	$\mathbf{h_{03} = (1.617E - 005)t^3 - 0.001t^2 + 1.71}$

6.3.2.3 The PHM by Including the Variable New Eras2

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
15800.370	809.586	2	.000	722.033	2	.000	722.033	2	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
New_Era2			593.114	2	.000			
New_Era2(1)	-4.326	.179	583.263	1	.000	.013	.009	.019
New_Era2(2)	-2.043	.133	236.973	1	.000	.130	.100	.168

The variables of the model were found to be statistically significant. It is time now to check if they satisfy the proportional hazard assumption by carrying out the Schoenfeld Residuals method of the SPSS software. Results of correlation of the partial residuals for NewEras2 are illustrated below

Correlations of partial residuals of NewEras2

		Partial residual for New_Era2(1)	Partial residual for New_Era2(2)	Rank of age
Partial residual for New_Era2(1)	Pearson Correlation	1	-.695**	.143**
	Sig. (2-tailed)		.000	.000
	N	1256	1256	1256
Partial residual for New_Era2(2)	Pearson Correlation	-.695**	1	-.087**
	Sig. (2-tailed)	.000		.002
	N	1256	1256	1256
Rank of age	Pearson Correlation	.143**	-.087**	1
	Sig. (2-tailed)	.000	.002	
	N	1256	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

The partial residuals were found to be significant for the variables NewEras2 which implies that the proportional hazard assumption is not satisfied, thus it is inappropriate to apply a PHM for this case.

The three variables will now be examined for possible statistical significance interaction. A separate PHM will be developed for each interaction term given that is statistically significant and satisfies the proportionality assumption.

6.3.2.4 The PHM for the Interaction Term (Length, NOPF)

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
16411.429	128.777	2	.000	110.974	2	.000	110.974	2	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.000	.000	24.062	1	.000	1.000	1.000	1.001
NOPF	.032	.004	69.195	1	.000	1.033	1.025	1.040

The variables in the model were found to be highly statistically significant. They have also been tested for the proportionality assumption and the resultant correlation matrix was

Correlations of partial residual for Length and NOPF

		Partial residual for L	Partial residual for NOPF	Rank of age
Partial residual for L	Pearson Correlation	1	.347**	.139**
	Sig. (2-tailed)		.000	.000
	N	1256	1256	1256
Partial residual for NOPF	Pearson Correlation	.347**	1	.031
	Sig. (2-tailed)	.000		.276
	N	1256	1256	1256
Rank of age	Pearson Correlation	.139**	.031	1
	Sig. (2-tailed)	.000	.276	
	N	1256	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

The test shows that the proportionality assumption is met for the variable NOPF but not for the variable Length. This implies that the model is not applicable for these two variables and a stratified model should be considered for the variable length.

The SCM by Length (L) adjusted for NOPF

Stratum Status[a]

Stratum	Strata label	Event	Censored	Censored Percent
1.00	0-500	844	401	32.2%
2.00	500-1000	319	87	21.4%
3.00	>1000	93	17	15.5%
Total		1256	505	28.7%

a. The strata variable is : LCatsR

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
14447.560	69.909	1	.000	63.406	1	.000	63.406	1	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
NOPF	.032	.004	68.698	1	.000	1.033	1.025	1.041

Correlations of partial residual for NOPF

		Partial residual for NOPF	Rank of age
Partial residual for NOPF	Pearson Correlation	1	.020
	Sig. (2-tailed)		.474
	N	1256	1256
Rank of age	Pearson Correlation	.020	1
	Sig. (2-tailed)	.474	
	N	1256	1256

The model was found to be statistically significant and it also satisfies the proportional hazard assumption. Graphs for the survival, one minus survival, log minus log, and hazard functions at mean for the SCM by length adjusted for NOPF are illustrated as follows

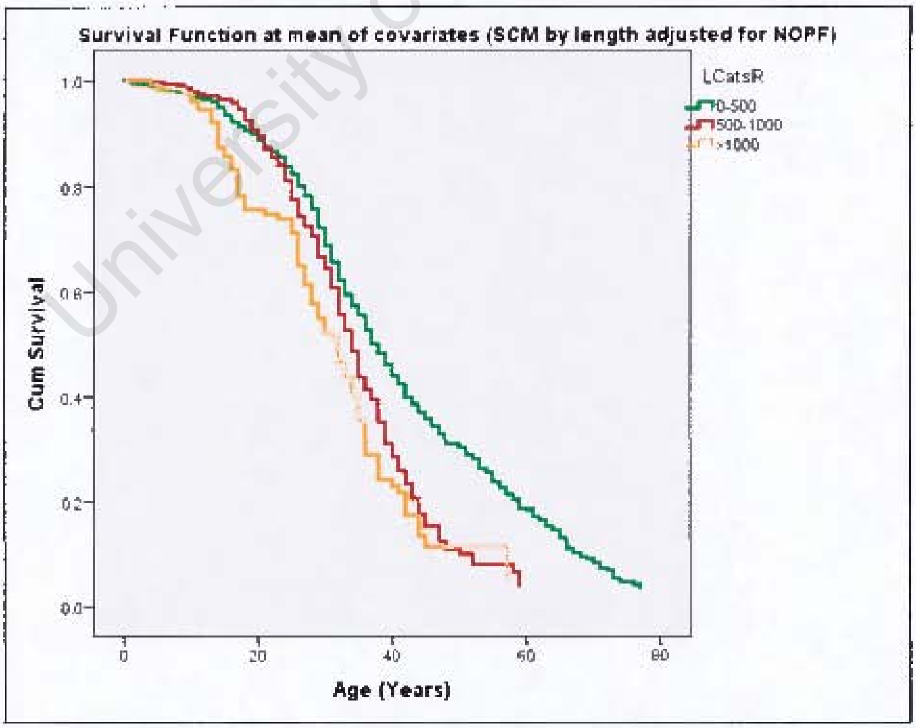


Figure 6-28: Survival functions (SCM by length adjusted for NOPF)

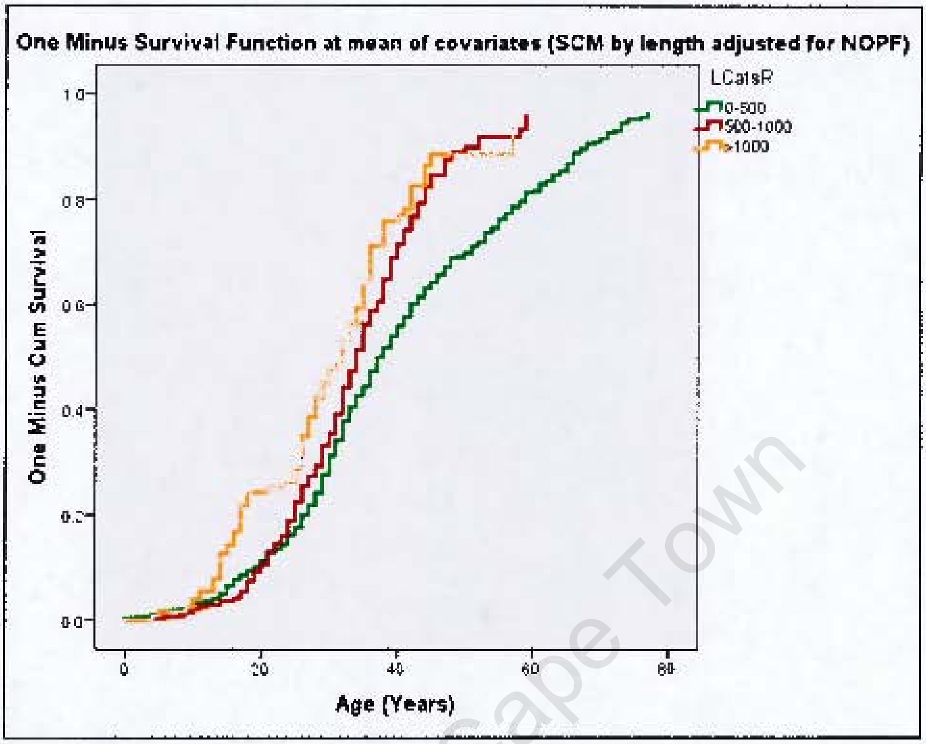


Figure 6-29: One minus survival functions (SCM by length adjusted for NOPF)

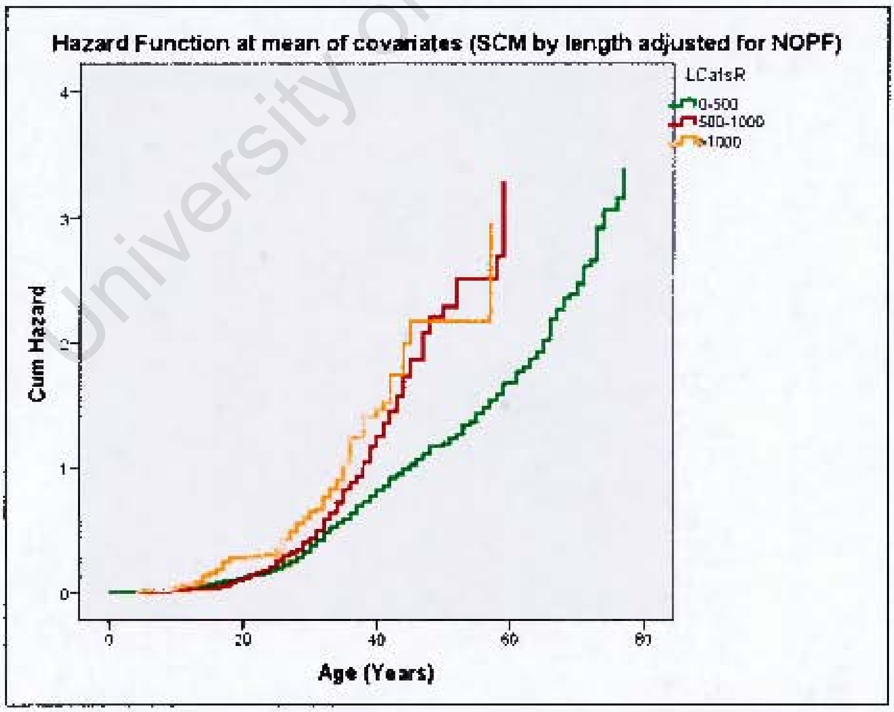


Figure 6-30: Hazard functions (SCM by length adjusted for NOPF)

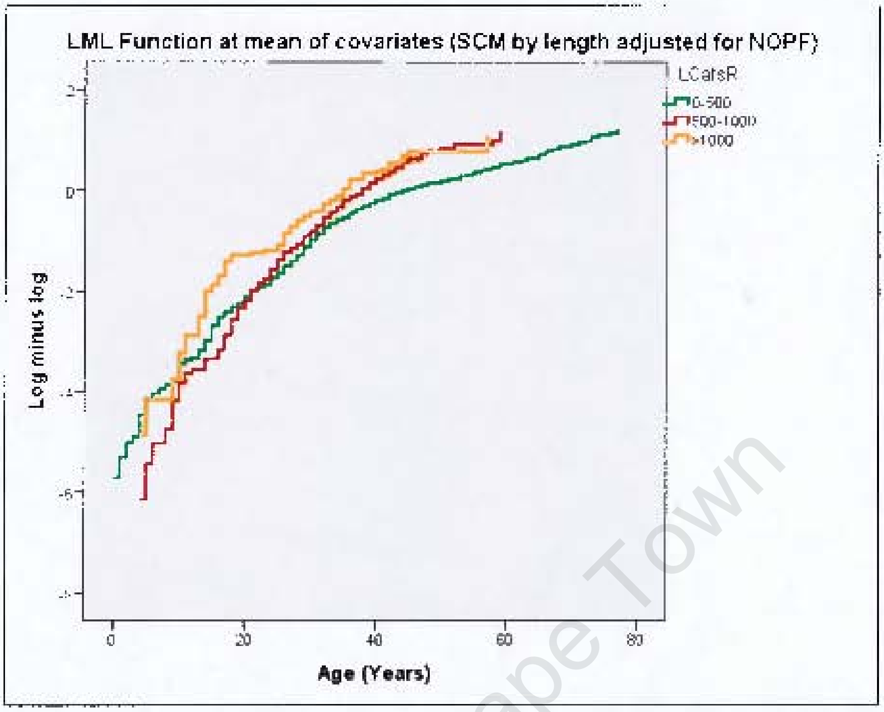


Figure 6-31: Log minus Log functions (SCM by length adjusted for NOPF)

Figure 6-28 clearly indicates that the probability of survival for pipes of length more than 1000 is less than the probability of survival for shorter pipes and the case remains the same for the other length categories. This result is consistent with earlier findings from previous studies which proved that longer pipes are more prone to failure than shorter pipes. Also from the results illustrated in Figure 6-30 is it evident that longer pipes encountered higher and earlier hazard rates. The hazard rate for pipes with length ranges from 0 to 500 km is almost half of the hazard rate for longer pipes. But this is to be expected since their longer length give them a high probability of failure, i.e. rate of failure/year×length. For full configuration of the PHM for these estimates for the baseline hazard function for the each stratum should be obtained.

Estimation of the baseline hazard functions for length strata

Baseline hazard function for LCatR1 ($0 < L < 500$)

Different models have been examined for goodness of fit and results are shown as

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.906	724.036	1	75	.000	-.434	.031		
Quadratic	.994	5687.654	2	74	.000	.027	-.006	.000483	
Cubic	.995	4467.944	3	73	.000	-.030	.004	.000177	2.669E-006
Compound	.906	725.978	1	75	.000	.015	1.081		
Growth	.906	725.978	1	75	.000	-4.226	.078		
Exponential	.906	725.978	1	75	.000	.015	.078		

The independent variable is AgeLcatrNOPF.

The cubic model was found to have a better R Square; however the quadratic model has been chosen for its better F value. The model summary is illustrated below

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AgeLcatrNOPF	-.005662	.001	-.173	-4.708	.000
AgeLcatrNOPF ** 2	.000483	.000	1.163	31.633	.000
(Constant)	.027352	.020		1.378	.172

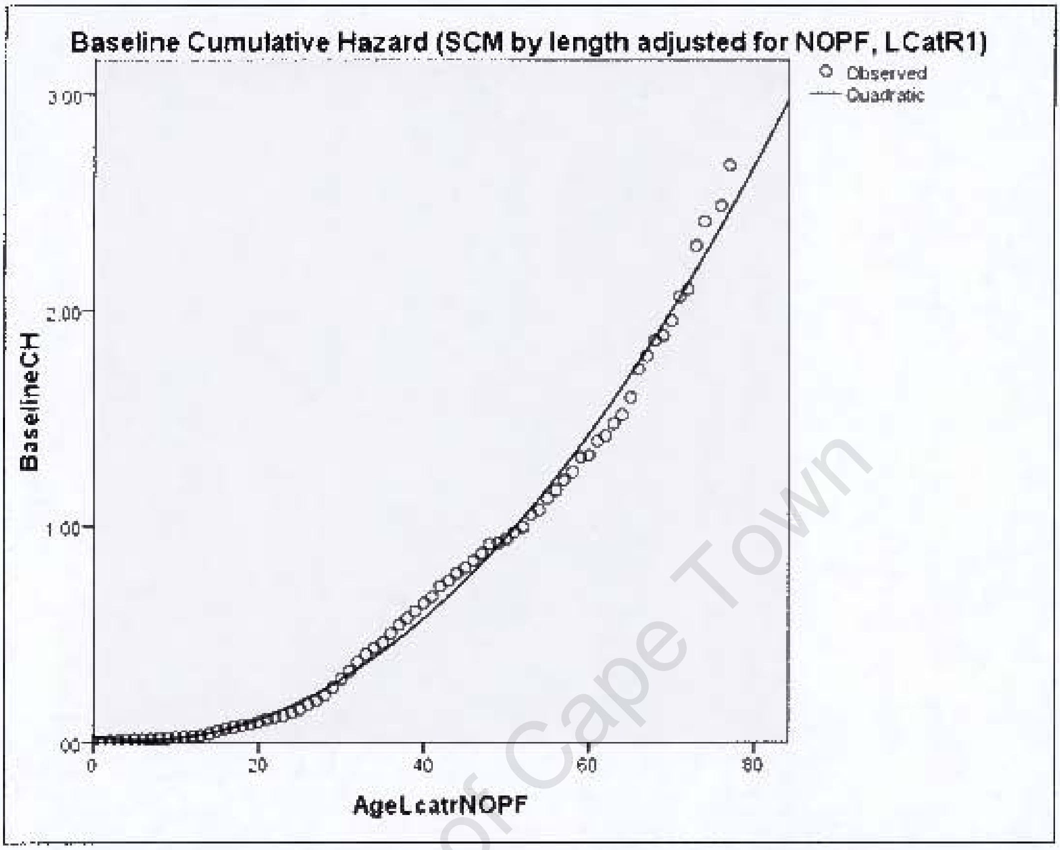


Figure 6-32: Baseline hazard function (SCM by length adjusted for NOPF, LCatR1)

Baseline hazard function for LCatR2 (500 < L < 1000)

The six models have been tested for goodness of fit for this length category and results are illustrated as

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH2

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.857	257.785	1	43	.000	-.650	.043		
Quadratic	.986	1445.757	2	42	.000	.086	-.021	.001	
Cubic	.989	1178.647	3	41	.000	.254	-.047	.002	-1.068E-005
Compound	.926	539.375	1	43	.000	.005	1.135		
Growth	.926	539.375	1	43	.000	-5.272	.127		
Exponential	.926	539.375	1	43	.000	.005	.127		

The independent variable is AgeLCatR2NOPF.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AgeLCatR2NOPF	-.021	.003	-.458	-6.221	.000
AgeLCatR2NOPF ** 2	.001	.000	1.429	19.426	.000
(Constant)	.086	.047		1.825	.075

Results show that the data is best fitted by the quadratic model.

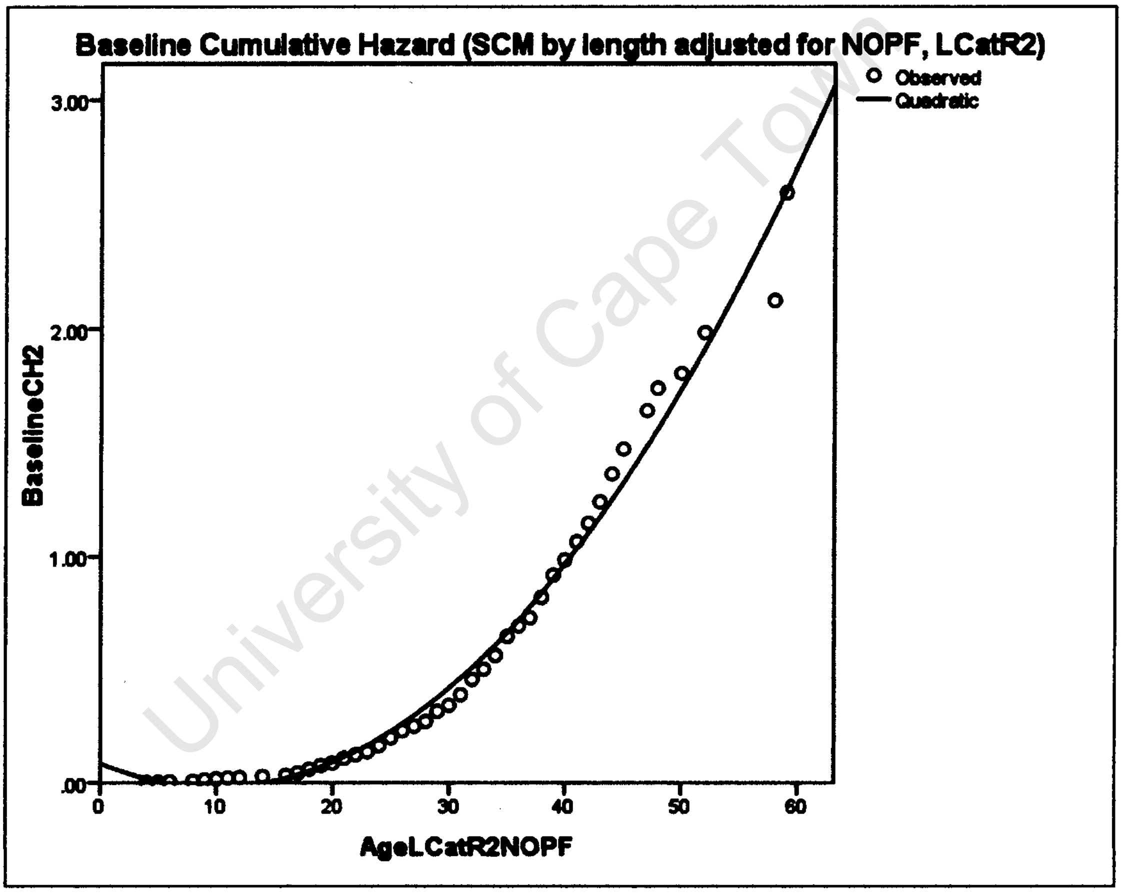


Figure 6-33: Baseline hazard function (SCM by length adjusted for NOPF, LCatR2)

Baseline bazard function for LCatR3 (L > 1000)

The data has again been tested for goodness of fit using different types of models and results revealed that the quadratic model is the most appropriated for curve fitting

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH3

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.875	210.863	1	30	.000	-.624	.042		
Quadratic	.980	697.735	2	29	.000	.040	-.011	.001	
Cubic	.983	546.108	3	28	.000	.705	-.037	.002	-1.130E-005
Compound	.899	267.582	1	30	.000	.015	1.116		
Growth	.899	267.582	1	30	.000	-4.173	.110		
Exponential	.899	267.582	1	30	.000	.015	.110		

The independent variable is AgeLCatR3NOPF.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AgeLCatR3NOPF	-.011	.004	-.251	-2.485	.019
AgeLCatR3NOPF ** 2	.001	.000	1.230	12.178	.000
(Constant)	.040	.058		.693	.494

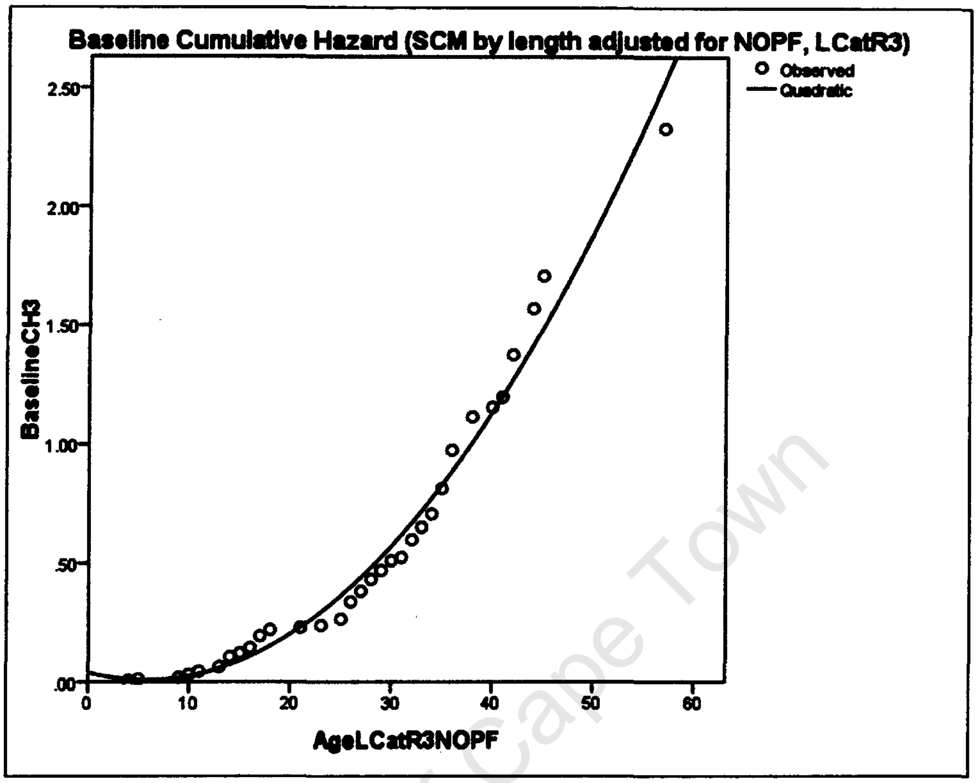


Figure 6-34: Baseline hazard function (SCM by length adjusted for NOPF, LCatR3)

The final PHM for strata of length with NOPF is the covariate is illustrated in Table 6-4.

Table 6-4: The SCM by Length adjusted for NOPF

Model Terms	Mathematical expression
The time independent term	$\exp(0.032 \times NOPF)$
Baseline hazard function (0 < L < 500 km): Quadratic	$h_{01} = 0.0005t^2 - 0.006t$
Baseline hazard function (500 < L < 1000 km): Quadratic	$h_{02} = 0.001t^2 - 0.021t + 0.086$
Baseline hazard function (L > 1000 km): Quadratic	$h_{03} = 0.001t^2 - 0.011t$

6.3.2.5 The PHM for the Interaction Term (Length, Neweras2)

Categorical Variable Codings[a]

		Frequency	(1)	(2)
New_Era2[b]	1=1930-1959	270	1	0
	2=1960-1989	1355	0	1
	3=1990-2010	137	0	0

a. Category variable: New_Era2 (Final New Eras)

b. Indicator Parameter Coding

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
15783.379	828.215	3	.000	739.025	3	.000	739.025	3	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.000	.000	18.095	1	.000	1.000	1.000	1.001
New_Era2			591.497	2	.000			
New_Era2(1)	-4.339	.179	585.259	1	.000	.013	.009	.019
New_Era2(2)	-2.102	.133	248.206	1	.000	.122	.094	.159

The three variables were found to be highly significant and the proportionality assumption has then been checked. Correlation was found to be insignificant for length whereas it was found to be significant for NewEras2 (1) and (2). That means the proportionality assumption is violated for these two variables and the PHM cannot be applied for this combination of variables. Alternatively, a SCM will be applied to strata of NewEras2 adjusted for L.

Correlations of partial residuals for Length and NewEras2

		Partial residual for L	Partial residual for New_Era2(1)	Partial residual for New_Era2(2)	Rank of age
Partial residual for L	Pearson Correlation	1	-.036	.118**	.042
	Sig. (2-tailed)		.204	.000	.138
	N	1256	1256	1256	1256
Partial residual for New_Era2(1)	Pearson Correlation	-.036	1	-.694**	.141**
	Sig. (2-tailed)	.204		.000	.000
	N	1256	1256	1256	1256
Partial residual for New_Era2(2)	Pearson Correlation	.118**	-.694**	1	-.085**
	Sig. (2-tailed)	.000	.000		.003
	N	1256	1256	1256	1256
Rank of age	Pearson Correlation	.042	.141**	-.085**	1
	Sig. (2-tailed)	.138	.000	.003	
	N	1256	1256	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

The SCM by NewEras2 adjusted for Length

Stratum Status[a]

Stratum	Strata label	Event	Censored	Censored Percent
1	1930-1959	198	72	26.7%
2	1960-1989	983	371	27.4%
3	1990-2010	75	62	45.3%
Total		1256	505	28.7%

a. The strata variable is : Final New Eras

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
15086.889	17.565	1	.000	16.498	1	.000	16.498	1	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.000342	.000	17.549	1	.000	1.000	1.000	1.001

Correlations

		Partial residual for L	Rank of age
Partial residual for L	Pearson Correlation	1	.047
	Sig. (2-tailed)		.094
	N	1256	1256
Rank of age	Pearson Correlation	.047	1
	Sig. (2-tailed)	.094	
	N	1256	1256

Results from applying a SCM to the dataset shows that the variable length (L) is highly significant; however it doesn't satisfy the proportional hazard assumption, therefore the model can't be applied for this case.

6.3.2.6 The PHM for the Interaction Term (NOPF, Neweras2)

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
15684.535	926.768	3	.000	837.869	3	.000	837.869	3	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
New_Era2			615.086	2	.000			
New_Era2(1)	-4.475	.181	608.408	1	.000	.011	.008	.016
New_Era2(2)	-2.119	.133	255.530	1	.000	.120	.093	.156
NOPF	.043	.004	132.123	1	.000	1.044	1.036	1.052

Correlations

		Partial residual for NOPF	Partial residual for New_Era2(1)	Partial residual for New_Era2(2)	Rank of age
Partial residual for NOPF	Pearson Correlation	1	.004	.042	.028
	Sig. (2-tailed)		.876	.141	.314
	N	1256	1256	1256	1256
Partial residual for New_Era2(1)	Pearson Correlation	.004	1	-.692**	.130**
	Sig. (2-tailed)	.876		.000	.000
	N	1256	1256	1256	1256
Partial residual for New_Era2(2)	Pearson Correlation	.042	-.692**	1	-.076**
	Sig. (2-tailed)	.141	.000		.007
	N	1256	1256	1256	1256
Rank of age	Pearson Correlation	.028	.130**	-.076**	1
	Sig. (2-tailed)	.314	.000	.007	
	N	1256	1256	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

The three variables were found to be highly statistically significant. Correlations of partial residual for the variables in the model showed that the proportionality assumption is not met for the variables NewEras2 which means that the model cannot be applied for this combination of variables. The model will therefore be constructed for groups of installation eras adjusted for NOPF.

The SCM by NewEras2 adjusted for NOPF

Stratum Status[a]

Stratum	Strata label	Event	Censored	Censored Percent
1	1930-1959	198	72	26.7%
2	1960-1989	983	371	27.4%
3	1990-2010	75	62	45.3%
Total		1256	505	28.7%

a. The strata variable is : Final New Eras

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
14992.185	129.751	1	.000	111.202	1	.000	111.202	1	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
NOPF	.042	.004	126.433	1	.000	1.043	1.035	1.051

Correlations of partial residual of NOPF

		Partial residual for NOPF	Rank of age
Partial residual for NOPF	Pearson Correlation	1	.011
	Sig. (2-tailed)		.709
	N	1256	1256
Rank of age	Pearson Correlation	.011	1
	Sig. (2-tailed)	.709	
	N	1256	1256

Graphs for the survival, one minus survival, log minus log, and hazard functions at mean for the SCM by NewEras2 adjusted for Length are illustrated as follows

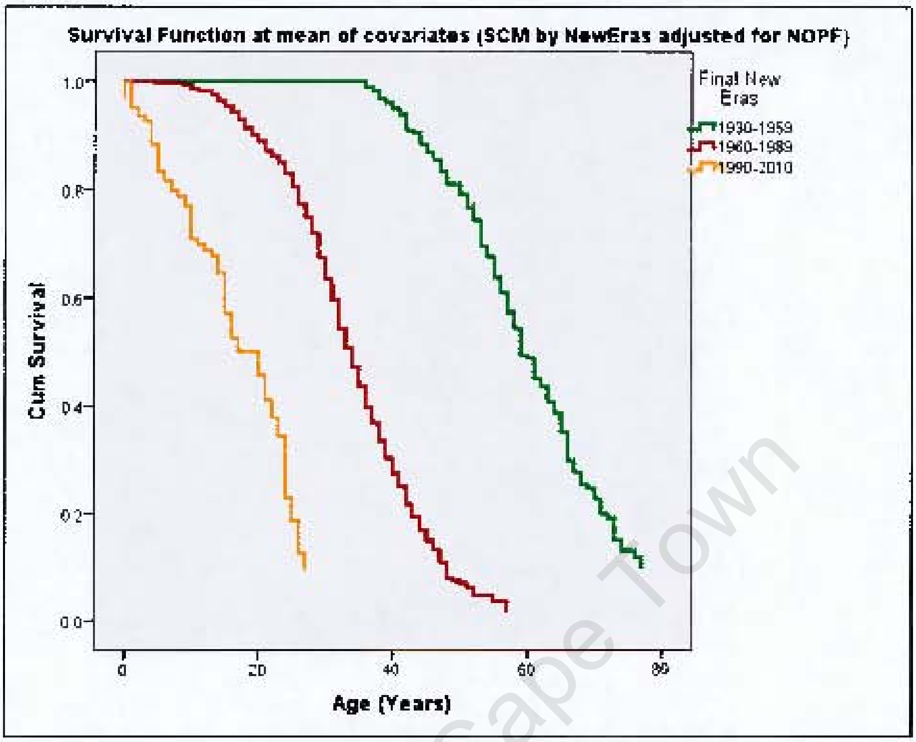


Figure 6-35: Survival function (SCM by NewEras2 adjusted for NOPF)

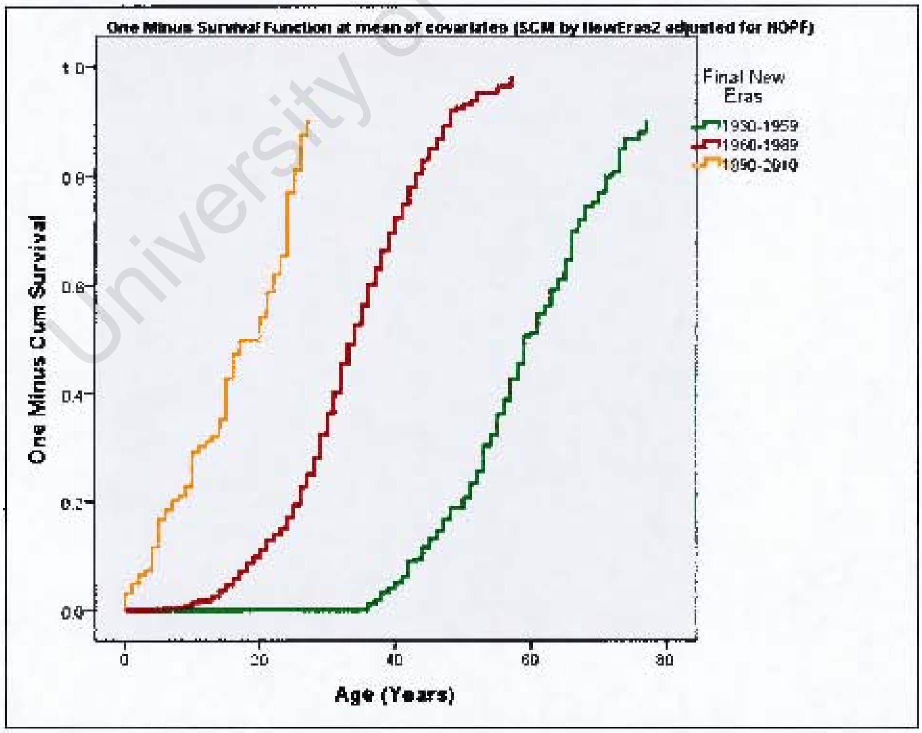


Figure 6-36: One minus survival function (SCM by NewEras2 adjusted for NOPF)

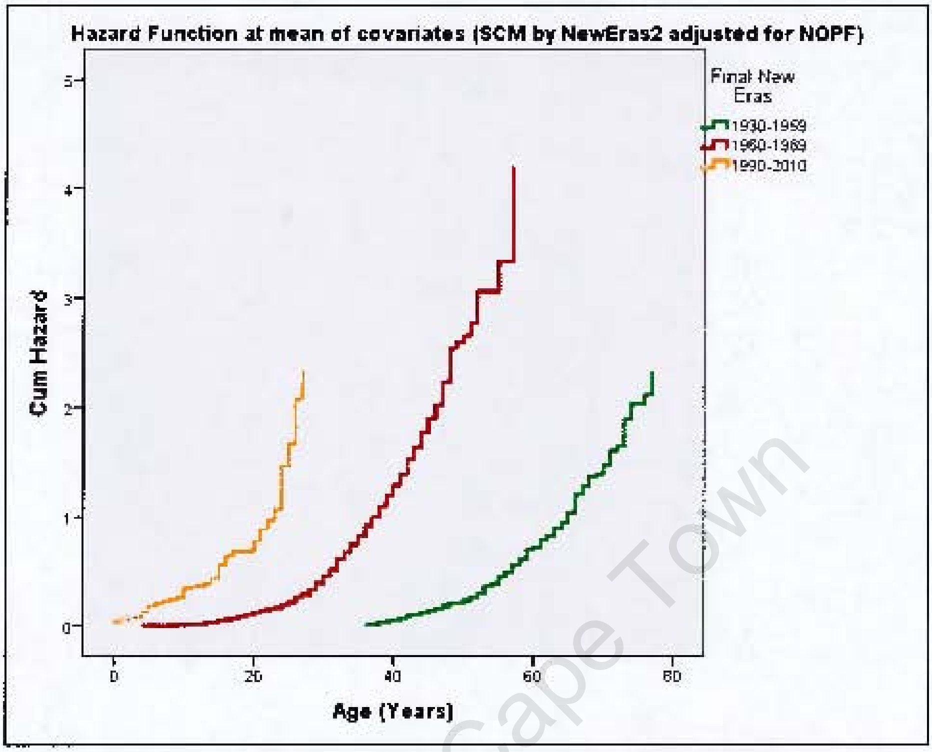


Figure 6-37: Hazard function (SCM by NewEras2 adjusted for NOPF)

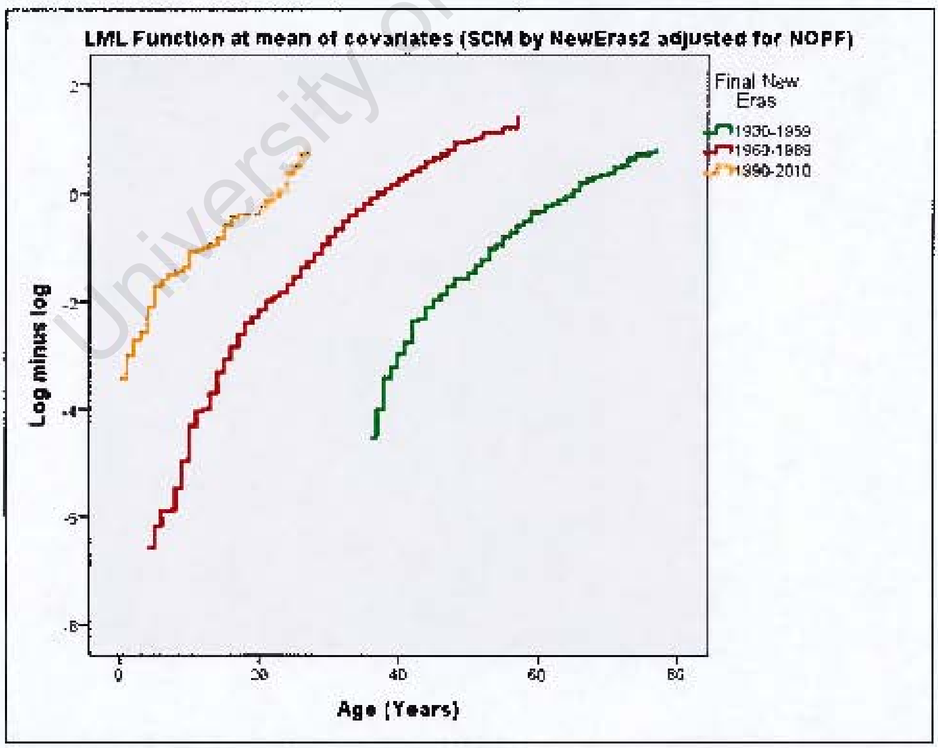


Figure 6-38: Log minus Log function (SCM by NewEras2 adjusted for NOPF)

Figure 6-35 reflects that the probability of survival for pipes installed between 1990 -2010 i.e. the third category, is far less than the probability of survival for pipes installed between 1960 - 1980 indicating some manufacturing defects in youngest pips. Constant probability of survival for pipes of the first installation age category for ages less than 40 years may indicate to the loss of failure records for this period of time.

Also, it is apparent from Figure 6-37 that pipes of the third installation period experience earlier wear out stage and higher hazard rates than pipes of the second installation era category which further indicates some defects in manufacturing quality. However, the hazard rate is dramatically increasing for pipes of the second installation eras for ages between 50 and 60 years indicating a rapid wear out of these pipes. Pipes of the first category seem to have normal wear out stage indicating good manufacturing standards and high reliability.

Final configuration of the SCM requires estimates for the baseline hazard functions for each stratum which is demonstrated as follows

Estimation of the baseline hazard functions for NewEras2 strata

Baseline hazard function for NewEra2 1(1930 - 1959)

Model Summary and Parameter Estimates

Dependent Variable: BCH1NewEra1NOPF

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.906	365.326	1	38	.000	-1.635	.038573		
Quadratic	.997	5779.938	2	37	.000	1.765	-.088403	.001131	
Cubic	.998	8034.449	2	37	.000	.198	.000000	-.000481	9.534E-006
Compound	.936	556.044	1	38	.000	.001	1.118348		
Growth	.936	556.044	1	38	.000	-7.580	.111853		
Exponential	.936	556.044	1	38	.000	.001	.111853		

The independent variable is Age1.

The cubic model was chosen to fit the data for its better properties than the quadratic model as shown from the models fitting results.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age1 ** 2	.000	.000	-1.339	-15.438	.000
Age1 ** 3	9.534E-006	.000	2.325	26.805	.000
(Constant)	.198	.032		6.110	.000

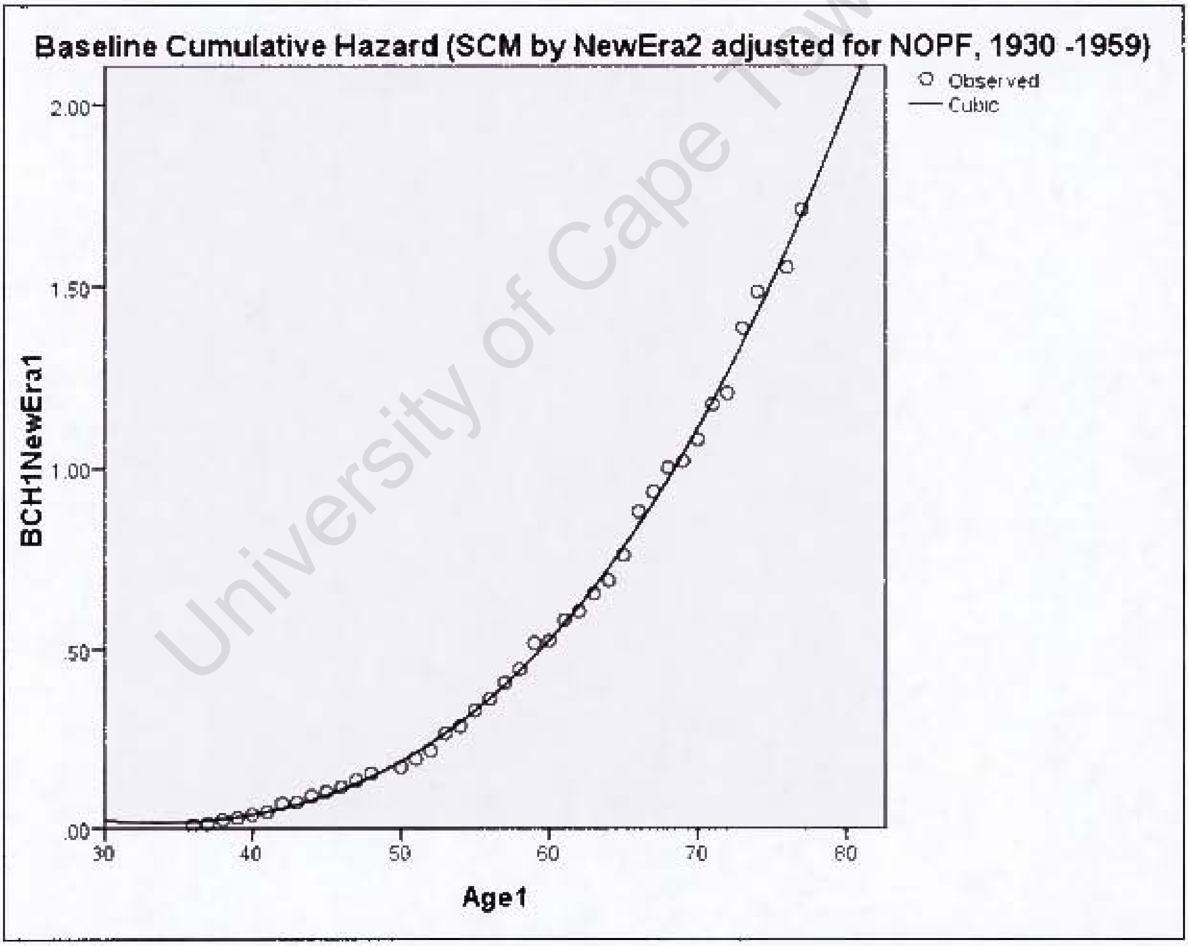


Figure 6-39: Baseline hazard function (SCM by NewEras2 adjusted for NOPF: 1930 - 1959)

Baseline hazard function for NewEra2 2 (1960 - 1989)

Model Summary and Parameter Estimates

Dependent Variable: BCH2NewEra2NOPF

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.825	226.419	1	48	.000	-.748	.048		
Quadratic	.993	3261.186	2	47	.000	.294	-.045	.002	
Cubic	.997	4824.687	3	46	.000	.044	-.006	3.638E-005	1.703E-005
Compound	.927	606.491	1	48	.000	.003	1.149		
Growth	.927	606.491	1	48	.000	-5.777	.139		
Exponential	.927	606.491	1	48	.000	.003	.139		

The independent variable is Age2.

The cubic model was found to better fit the data so it has been used for model development for this installation era.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age2	-.006	.005	-.117	-1.125	.266
Age2 ** 2	3.638E-005	.000	.042	.179	.859
Age2 ** 3	1.703E-005	.000	1.064	7.608	.000
(Constant)	.044	.042		1.065	.293

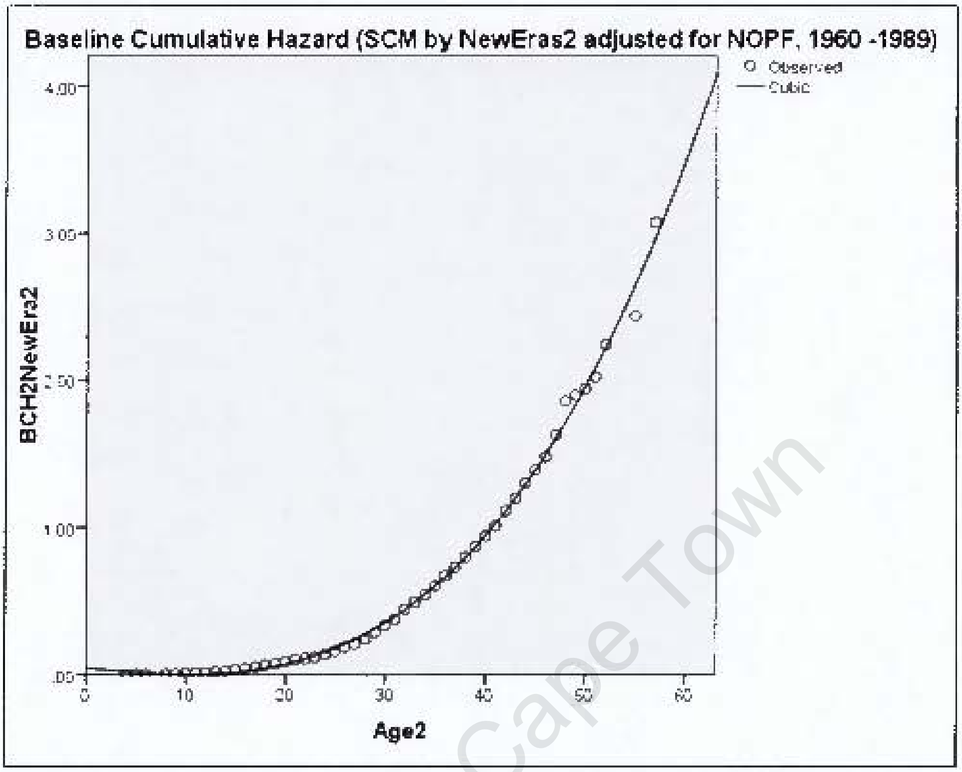


Figure 6-40: Baseline hazard function (SCM by NewEras2 adjusted for NOPF: 1960-1989)

Baseline hazard function for NewEra2 3 (1990 - 2010)

Model Summary and Parameter Estimates

Dependent Variable: BCH3NewEra3NOPF

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.816	106.579	1	24	.000	-.186	.050		
Quadratic	.941	184.824	2	23	.000	.131	-.024	.003	
Cubic	.981	380.024	3	22	.000	-.053	.06849	-.00624	.000223
Compound	.949	451.064	1	24	.000	.048	1.142		
Growth	.949	451.064	1	24	.000	-3.031	.133		
Exponential	.949	451.064	1	24	.000	.048	.133		

The independent variable is Age3.

Results from fitting the data to the six curves showed that the cubic model is the best for this installation era

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age3	.06849	.015	1.247	4.556	.000
Age3 ** 2	-.00624	.001	-3.201	-4.667	.000
Age3 ** 3	.000223	.000	2.996	6.788	.000
(Constant)	-.053	.045		-1.174	.253

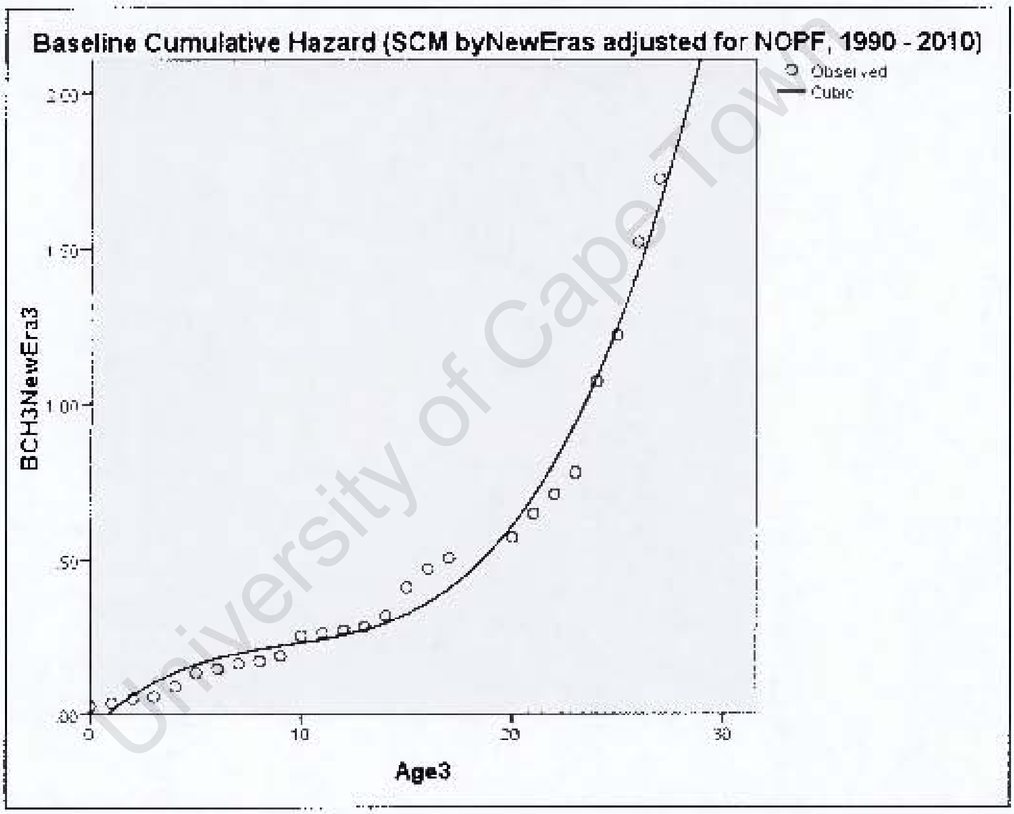


Figure 6-41: Baseline hazard function (SCM by NewEras2 adjusted for NOPF: 1990 - 2010)

The final SCM f by NewEras2 adjusted for NOPF is shown in Table 6-5.

Table 6-5: The SCM by NewEras2 adjusted for NOPF

Model Terms	Mathematical expression
The time independent term	$\exp(.042 \times NOPF)$
Baseline hazard function (1930-1959): Cubic	$h_{01} = (9.534E-006)t^3 - 0.000481t^2 + 0.198$
Baseline hazard function (1960-1989): Cubic	$h_{02} = (1.703E-005)t^3 + 3.64E-005t^2 - 0.006t + 0.044$
Baseline hazard function (1990 -2010): Cubic	$h_{03} = 0.000223t^3 - 0.006t^2 + .068t - 0.053$

6.3.2.7 The PHM for the Interaction Term (L, NOPF, Neweras2)

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
15684.454	930.289	4	.000	837.950	4	.000	837.950	4	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.000	.000	.081	1	.775	1.000	1.000	1.000
NOPF	.042	.004	108.943	1	.000	1.043	1.035	1.052
New_Era2			612.716	2	.000			
New_Era2(1)	-4.474	.182	607.187	1	.000	.011	.008	.016
New_Era2(2)	-2.123	.133	254.318	1	.000	.120	.092	.156

Results from running the PHM for this combination of factors showed that the length L is not significant. The model has been applied to the same combination with transforming the variable Length into LnL (the natural logarithm of L) and Sqrt L (the square root of L) and they were all found to be insignificant. Other combinations have been examined for the three variables by stratifying the variable length but they were all violated in that they don't satisfy the PHM requirements. It was therefore recommended to examine the time effect of the variable Length by applying the Extended Cox Model to further investigate the multiplicative effect of the three variables in the hazard of failure of water pipelines.

Correlations

		Partial residual for L	Partial residual for NOPF	Partial residual for New_Era2(1)	Partial residual for New_Era2(2)	Rank of age
Partial residual for L	Pearson Correlation	1	.345**	-.030	.114**	.067*
	Sig. (2-tailed)		.000	.280	.000	.018
	N	1256	1256	1256	1256	1256
Partial residual for NOPF	Pearson Correlation	.345**	1	.005	.041	.030
	Sig. (2-tailed)	.000		.868	.143	.290
	N	1256	1256	1256	1256	1256
Partial residual for New_Era2(1)	Pearson Correlation	-.030	.005	1	-.692**	.130**
	Sig. (2-tailed)	.280	.868		.000	.000
	N	1256	1256	1256	1256	1256
Partial residual for New_Era2(2)	Pearson Correlation	.114**	.041	-.692**	1	-.076**
	Sig. (2-tailed)	.000	.143	.000		.007
	N	1256	1256	1256	1256	1256
Rank of age	Pearson Correlation	.067*	.030	.130**	-.076**	1
	Sig. (2-tailed)	.018	.290	.000	.007	
	N	1256	1256	1256	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

It is apparent, from the previous analysis, that the number of previous failures (NOPF) is an important factor that affects the hazard rate of pipelines failures. It might, therefore, be useful to explore the failure trends of pipes of different categories of NOPF as well as the effect of the variable NOPF as a discrete variable. A PHM has thus been applied by including the variable NewNOPF (a discrete variable) and a SCM by including the variable

NOPF (as a continuous variable) will was also applied adjusted for the two other variables. Results are presented in the next section.

6.3.2.8 PHM at Pipe Level by including NewNOPF as dummy variable

The model has been examined by including variable is NewNOPF and also for interaction term between the three variables. Results are illustrated as

Categorical Variable Codings[a]

		Frequency	(1)	(2)	(3)
New_NoOfPF[b]	1=1-4	731	1	0	0
	2=5-10	648	0	1	0
	3=11-20	260	0	0	1
	4=>20	123	0	0	0

a. Category variable: New_NoOfPF (New No of Prev Failures)

b. Indicator Parameter Coding

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
16357.776	165.811	3	.000	164.628	3	.000	164.628	3	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
NewNOPF			157.563	3	.000			
NewNOPF(1)	-.629	.107	34.496	1	.000	.533	.432	.658
NewNOPF(2)	.016	.102	.023	1	.879	1.016	.831	1.241
NewNOPF(3)	.374	.114	10.804	1	.001	1.454	1.163	1.817

Results from applying the PHM with the NewNOPF is the covariate shows that the third category of NOPF is not statistically significant. The variable has therefore been tested for any interaction trends with the other two variables. Results are illustrated as

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
15657 721	951.880	6	.000	864.682	6	.000	864 682	6	000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.000	.000	3.174	1	.075	1 000	1 000	1.000
NewNOPF			117.465	3	.000			
NewNOPF (1)	.792	.110	51.899	1	.000	.453	.365	.562
NewNOPF(2)	-.182	.105	3.018	1	.082	.833	.678	1.024
NewNOPF(3)	-.003	.116	.001	1	.978	.997	.795	1.250
New_Era2			569.137	2	.000			
New_Era2(1)	-4.304	.181	562.708	1	.000	.014	.009	.019
New_Era2(2)	-2.049	.134	234 701	1	.000	.129	.099	.167

Results from applying the PHM by including the variable NOPF as a dummy variable; shows that the first and second categories of NOPF are highly statistically significant whereas the other two categories are not. It is apparent from the previous results that the statistical significance of the model could be maximized if the variable NOPF is included as a continuous variable. To investigate any trends of failure among pipes of different categories of NOPF a SCM has been applied for this variable adjusted for the other two variables.

6.3.2.9 The SCM by NOPF adjusted for (L, Neweras2)

Stratum Status[a]

Stratum	Strata label	Event	Censored	Censored Percent
1	1-4	347	384	52.5%
2	5-10	551	97	15.0%
3	11-20	240	20	7.7%
4	>20	118	5	4.1%
Total		1256	506	28.7%

a. The strata variable is : New No of Prev Failures

Categorical Variable Codings[a]

		Frequency	(1)	(2)
New_Era2[b]	1=1930-1959	270	1	0
	2=1960-1989	1355	0	1
	3=1990-2010	137	0	0

a. Category variable: New_Era2 (Final New Eras)

b. Indicator Parameter Coding

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
12577.087	717.933	3	.000	673.663	3	.000	673.663	3	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
L	.000	.000	2.209	1	.137	1.000	1.000	1.000
New_Era2			535.809	2	.000			
New_Era2(1)	-4.193	.183	525.705	1	.000	.015	.011	.022
New_Era2(2)	-1.980	.137	209.521	1	.000	.138	.106	.181

The variable length was found to insignificant so the mode has been applied by including the variable NewEras2 only.

6.3.2.10The SCM BY NOPF adjusted for (Neweras2)

Stratum Status[a]

Stratum	Strata label	Event	Censored	Censored Percent
1	1-4	347	384	52.5%
2	5-10	551	97	15.0%
3	11-20	240	20	7.7%
4	>20	118	5	4.1%
Total		1256	506	28.7%

a. The strata variable is : New No of Prev Failures

Categorical Variable Codings[a]

		Frequency	(1)	(2)
New_Era2[b]	1=1930-1959	270	1	0
	2=1960-1989	1355	0	1
	3=1990-2010	137	0	0

a. Category variable. New_Era2 (Final New Eras)

b. Indicator Parameter Coding

Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
12579.252	715.771	2	.000	671.498	2	.000	671.498	2	.000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
New_Era2			537.649	2	.000			
New_Era2(1)	-4.190	.183	525.928	1	.000	.015	.011	.022
New_Era2(2)	-1.957	.136	207.707	1	.000	.141	.108	.184

Correlations of partial residuals

		Partial residual for New_Era2(1)	Partial residual for New_Era2(2)	Rank of age
Partial residual for New_Era2(1)	Pearson Correlation	1	-.702**	.147**
	Sig. (2-tailed)		.000	.000
	N	1256	1256	1256
Partial residual for New_Era2(2)	Pearson Correlation	-.702**	1	-.090**
	Sig. (2-tailed)	.000		.001
	N	1256	1256	1256
Rank of age	Pearson Correlation	.147**	-.090**	1
	Sig. (2-tailed)	.000	.001	
	N	1256	1256	1256

**. Correlation is significant at the 0.01 level (2-tailed).

The test for partial residual for the explanatory variable NewEras2 shows that the model satisfies the proportionality assumption. Curves for survival, one minus survival, hazard and log minus log function are illustrated in the figures below

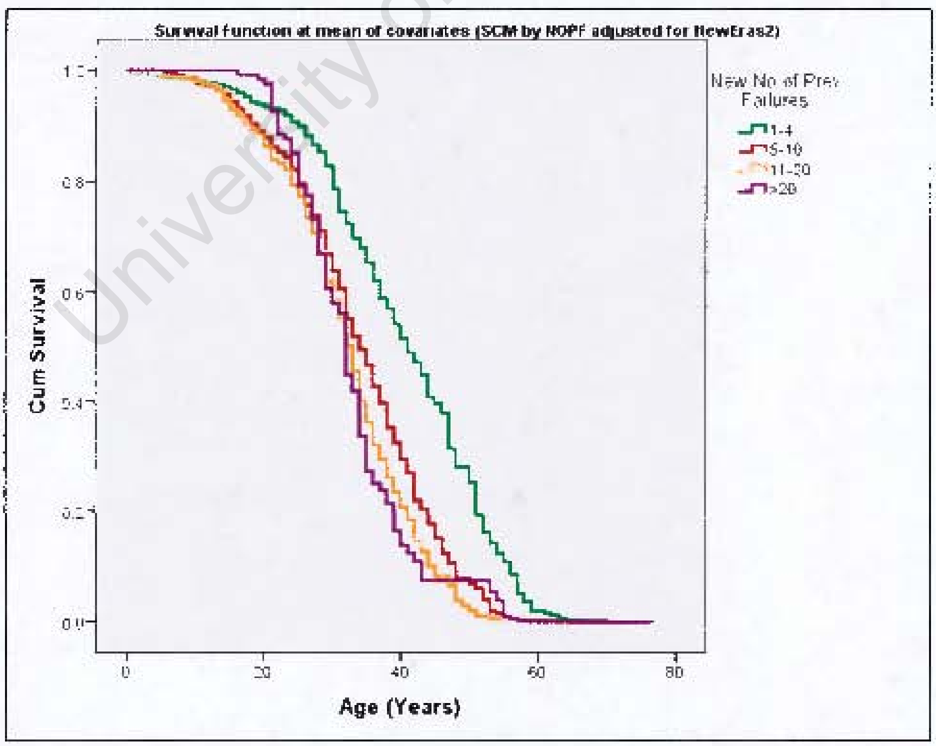


Figure 6-42: Survival function (SCM by NOPF adjusted for NewEras2)

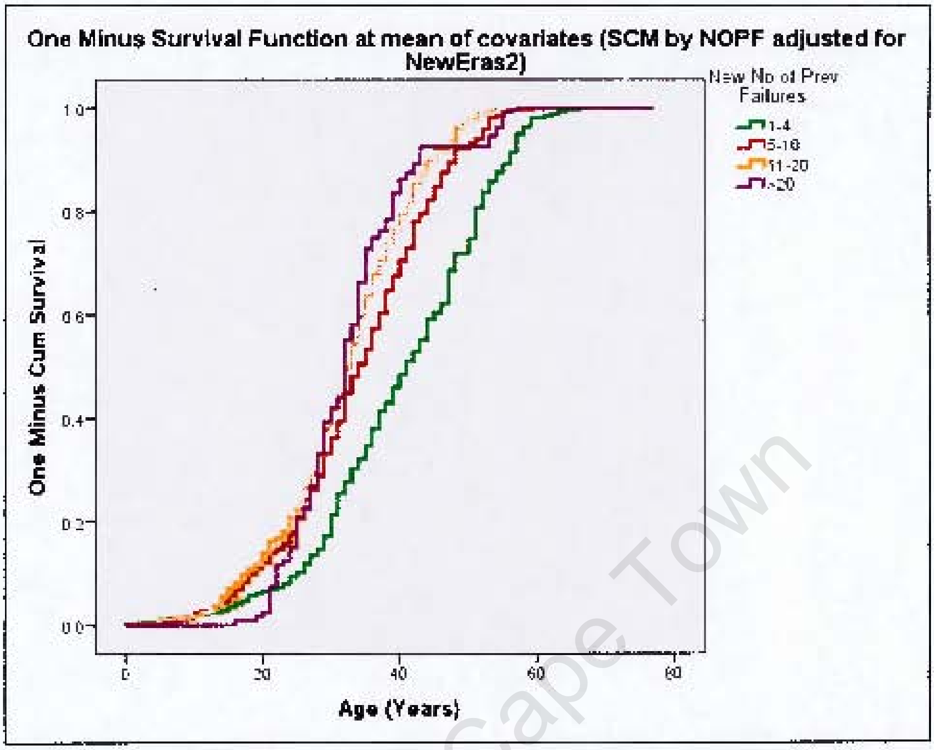


Figure 6-43: One minus survival function (SCM by NOPF adjusted for NewEras2)

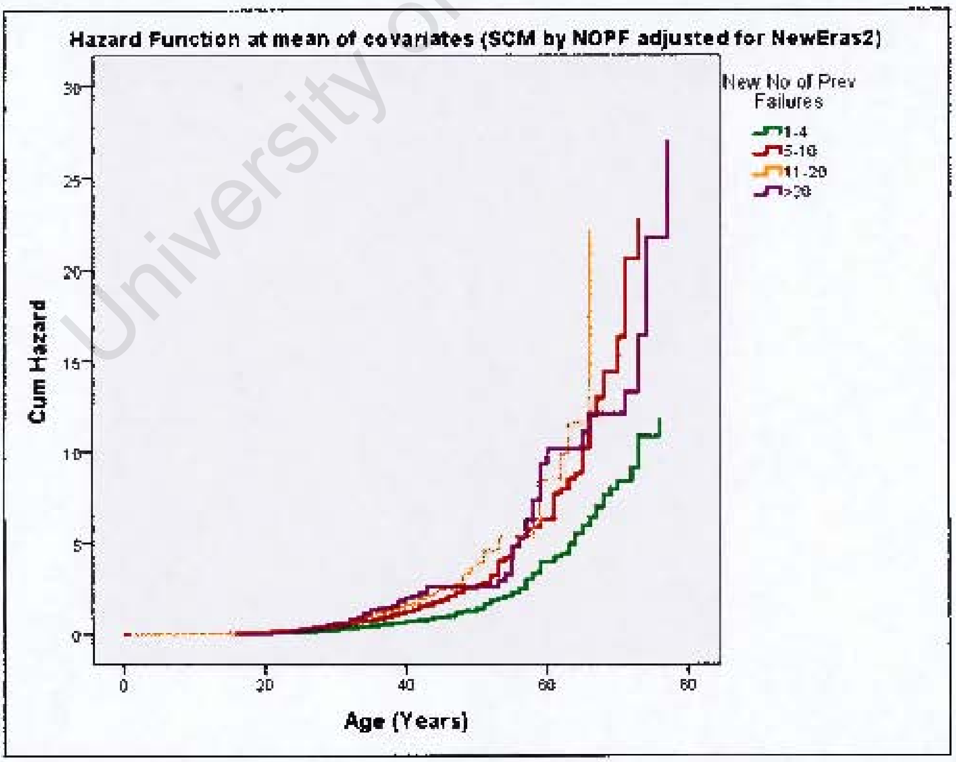


Figure 6-44: Hazard function (SCM by NOPF adjusted for NewEras2)

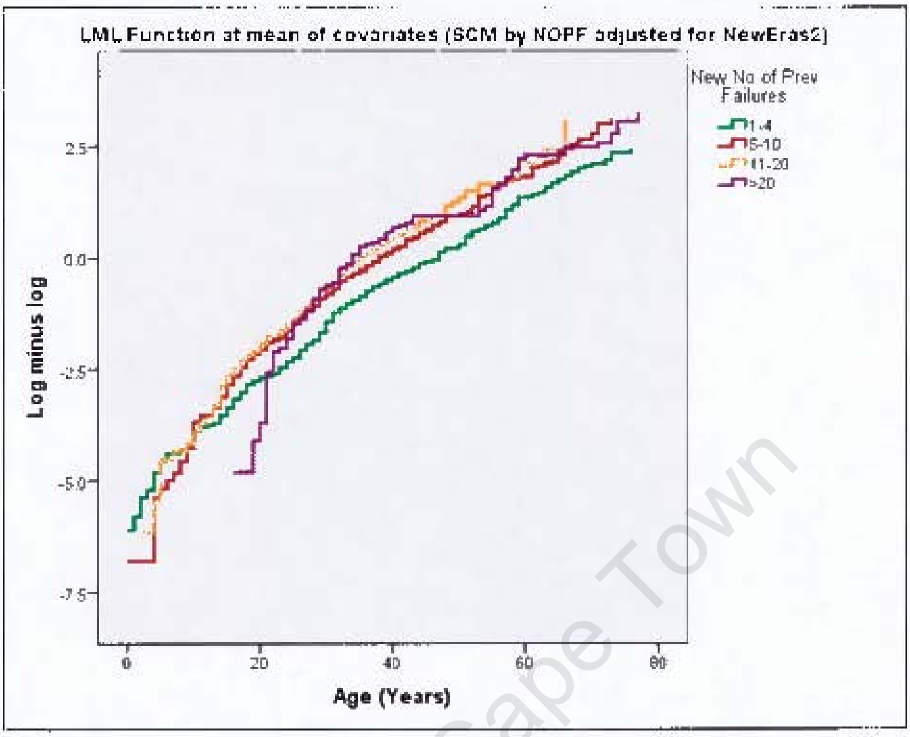


Figure 6-45: Log minus Log function (SCM by NOPF adjusted for NewEras2)

Figure 6-42, shows that higher NOPF is associated with lower probability of survival, as clearly shown for the first three categories of NOPF, with normal drop in the probability of survival. Pipes of the 4th category of NOFP (i.e. pipes with more than 20 failures) are experiencing a slightly different trend of failure. The probability of survival for this category of pipes remains approximately constant (i.e. 1) up to age 20 years, where it drops dramatically until it reaches (0.18) for pipes of age 40 years. It then remains constant for ages (42-55 years) where it starts to drop again to the minimum probability of survival. Conversely, as it is clear from Figure 6-44, that the hazard rate increases as the NOPF increases for the first three categories of NOPF. The hazard rate for pipes of the 4th category remains increasing with the same trend of the other three categories until age 42 years where it becomes constant until 55 years and it starts to increase dramatically from (3) at age 42 years to (27) at age 78 years with high ties indicating frequent occurrences of failure at this period.

Full configuration of the SCM requires the estimation of baseline hazard function for the four categories of NOPF which are estimated as follows

Estimation of the baseline hazard functions for NOPF strata

Baseline hazard function for NOPF 1(1 - 4)

The previously defined six models have again been tested for goodness of fit for the data of this category. Results are illustrated as

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH1

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.662	134.981	1	69	.000	-17.892	.926		
Quadratic	.944	572.596	2	68	.000	9.951	-1.339	.031	
Cubic	.995	4108.867	3	67	.000	-3.047	.832	-.043	.001
Compound	.983	3929.503	1	69	.000	.059	1.112		
Growth	.983	3929.503	1	69	.000	-2.832	.106		
Exponential	.983	3929.503	1	69	.000	.059	.106		

The independent variable is AgeNOPF1.

Although the compound, growth and exponential modes produced the highest R Square, the cubic model has been chosen for it better F value.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AgeNOPF1	.832	.095	.732	8.734	.000
AgeNOPF1 ** 2	-.043	.003	-2.910	-14.447	.000
AgeNOPF1 ** 3	.001	.000	3.172	25.053	.000
(Constant)	-3.047	.823		-3.701	.000

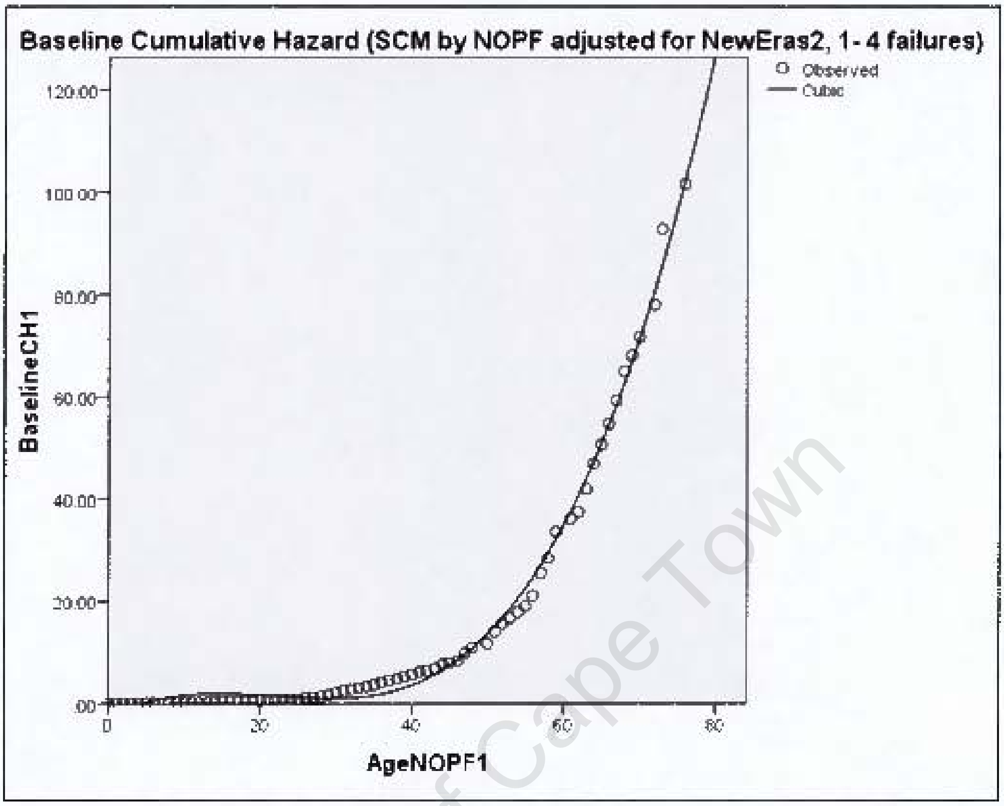


Figure 6-46: Baseline hazard function (SCM by NOPF adjusted for NewEras2)

Baseline hazard function for NOPF (5 - 10)

Results from model fitting are

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH2

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig	Constant	b1	b2	b3
Linear	.631	109.590	1	64	.000	-35.120	1.705		
Quadratic	.910	316.737	2	63	.000	23.913	-2.782	.060	
Cubic	.979	947.176	3	62	.000	-12.849	2.612	-.119	.002
Compound	.958	1447.929	1	64	.000	.073	1.123		
Growth	.958	1447.929	1	64	.000	-2.620	.116		
Exponential	.958	1447.929	1	64	.000	.073	.116		

The independent variable is AgeNOPF2.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AgeNOPF2	.116	.003	.979	38.052	.000
(Constant)	.073	.009		7.836	.000

The dependent variable is ln(BaselineCH2).

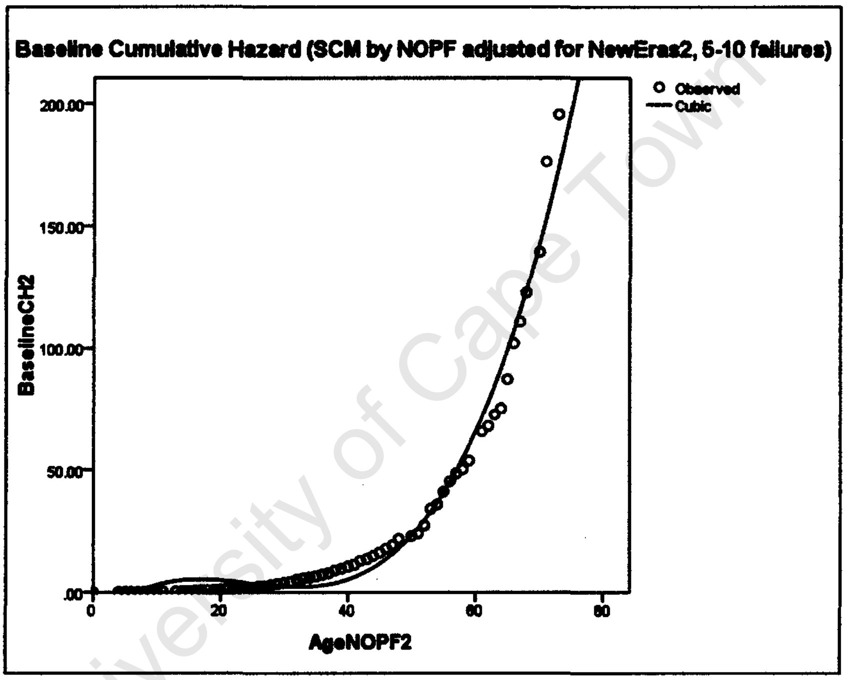


Figure 6-47: Baseline hazard function (SCM by NOPF adjusted for NewEras2, The Cubic Model)

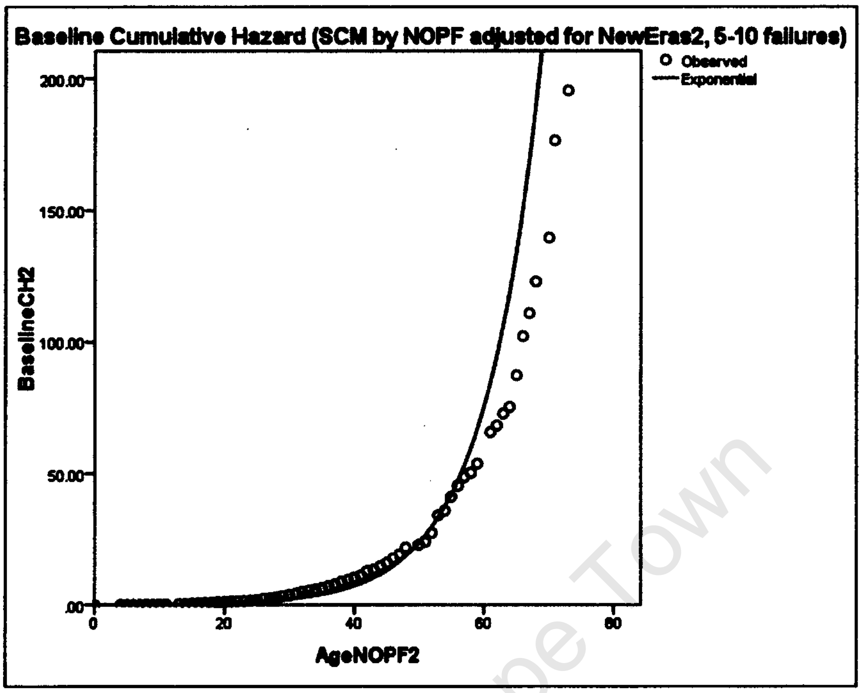


Figure 6-48: Baseline hazard function (SCM by NOPF adjusted for NewEras2, The Exp. Model)

Results from fitting the data to the six curve suggests that the cubic, compound, growth and exponential curves are the best for this category of NOPF. However, the exponential model may be preferred due to its better F value. It is also apparent, from Figure 6-47 and Figure 6-48, that the data may be better represented by two different models, one for ages until 50 years and the other for older pipes.

Baseline hazard function for NOPF (11 - 20)

Results from examining the six curves for goodness of fit are

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH3

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.549	60.824	1	50	.000	-27.286	1.438		
Quadratic	.845	133.125	2	49	.000	20.228	-2.480	.060	
Cubic	.924	193.481	3	48	.000	-11.929	2.465	-.119	.002
Compound	.959	1183.722	1	50	.000	.073	1.134		
Growth	.959	1183.722	1	50	.000	-2.621	.126		
Exponential	.959	1183.722	1	50	.000	.073	.126		

The independent variable is AgeNOPF3.

The exponential model may also be preferred for fitting the data for this category of NOPF. It was also noticed that the baseline hazard function may be better represented by two curves rather than one model over time. Another model may be used to fit the data from age 50 years on.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig
	B	Std. Error	Beta		
AgeNOPF3	.126	.004	.980	34.405	.000
(Constant)	.073	.009		7.743	.000

The dependent variable is ln(BaselineCH3).

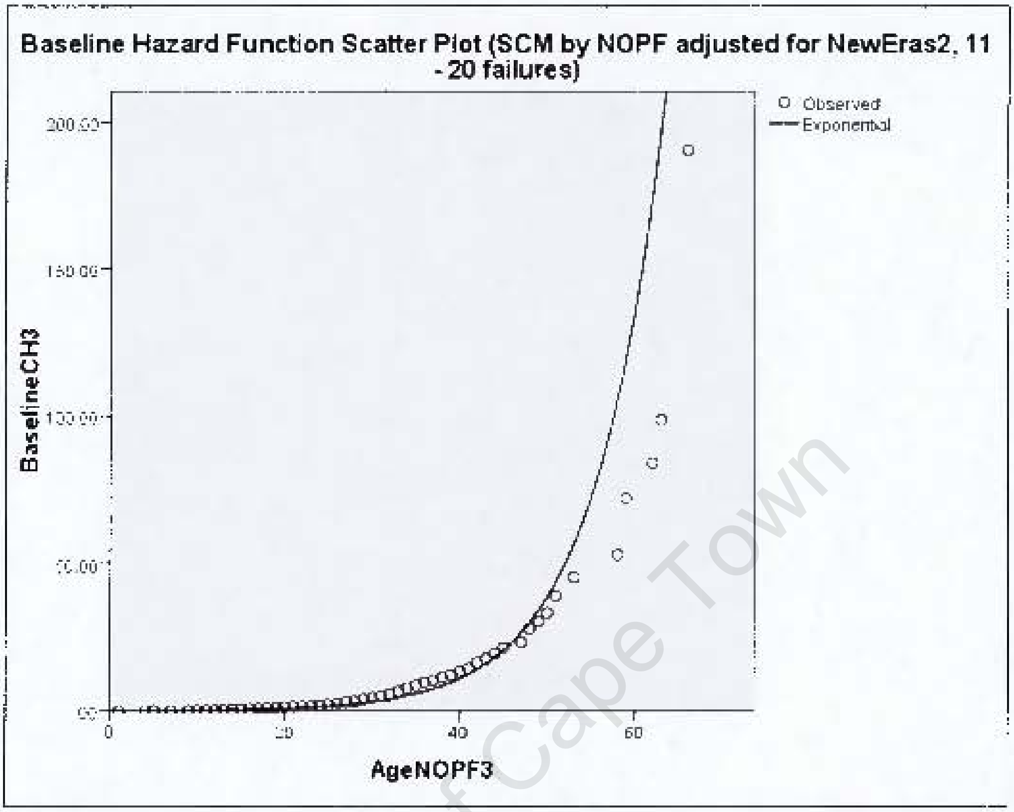


Figure 6-49: Baseline hazard function (SCM by NOPF adjusted for NewEras2, 11 – 20 failures)

Baseline hazard function for NOPF > 20

Model Summary and Parameter Estimates

Dependent Variable: BaselineCH4

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.771	127.991	1	38	.000	-76.476	2.722		
Quadratic	.944	309.957	2	37	.000	72.170	-4.900	.083	
Cubic	.969	375.622	3	36	.000	-65.274	6.012	-.176	.002
Compound	.871	256.733	1	38	.000	.123	1.112		
Growth	.871	256.733	1	38	.000	-2.096	.106		
Exponential	.871	256.733	1	38	.000	.123	.106		

The independent variable is AgeNOPF4.

The cubic model was found to be the better in representing the data for this category of NOPF.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
AgeNOPF4	6.012	2.082	1.939	2.888	.007
AgeNOPF4 ** 2	-.176	.048	-5.283	-3.662	.001
AgeNOPF4 ** 3	.002	.000	4.349	5.431	.000
(Constant)	-65.274	27.701		-2.356	.024

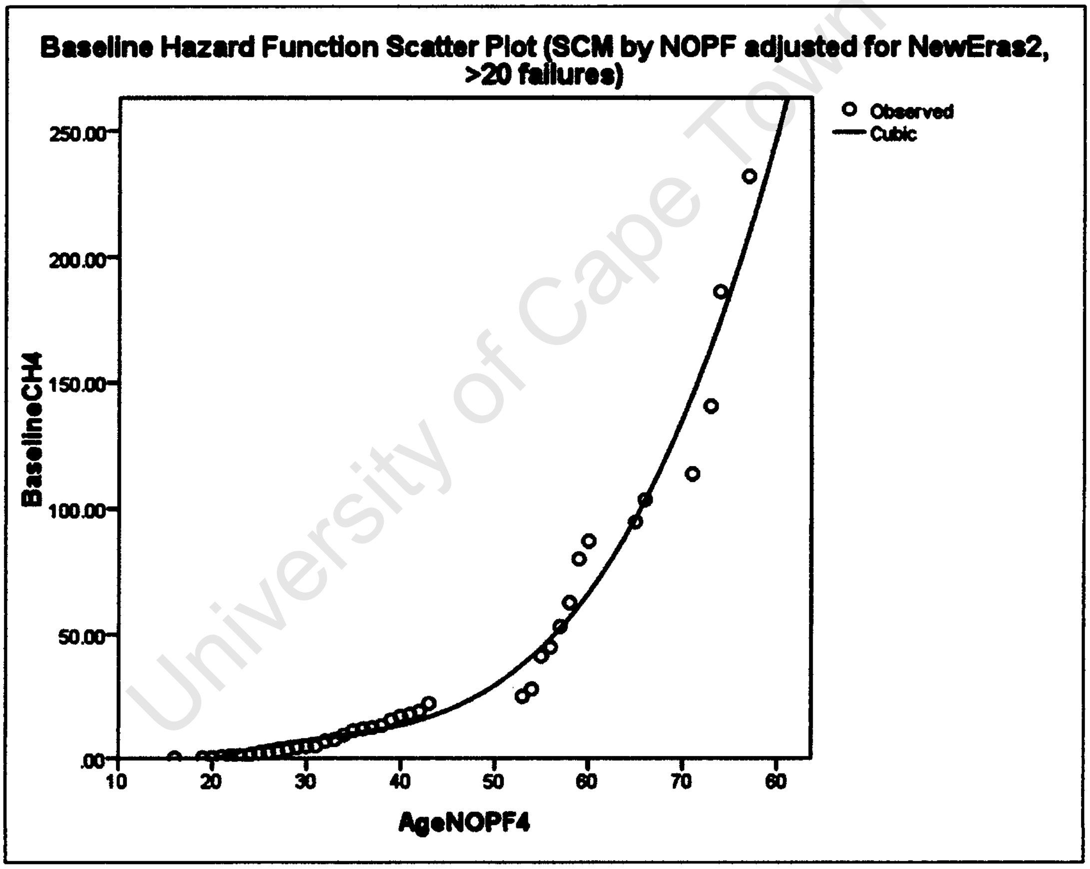


Figure 6-50: Baseline hazard function (SCM by NOPF adjusted for NewEras2, > 20 failures)

The final SCM f by NOPF adjusted for NewEras2 is shown in Table 6-6.

Table 6-6: The SCM by NOPF adjusted for NewEras2

Model Terms	**Mathematical expression**
The time independent term	$\exp(-4.190 \times \text{NewEra2}(1) - 1.957 \times \text{NewEras2}(2))$
Baseline hazard function (1-4): Cubic	$h_{01} = 0.001t^3 - 0.043\,t^2 + 0.832\,t - 3.047$
Baseline hazard function (5-10): Exponential	$\ln h_{02} = 0.116\,t + 0.073$
Baseline hazard function (11-20): Exponential	$\ln h_{03} = 0.126\,t + 0.073$
Baseline hazard function (>20): Cubic	$h_{04} = 0.002\,t^3 - 0.176\,t^2 + 6.012\,\text{t} - 65.274$

6.3.2.11 Discussion

In section 6.3.2, the PHM has been applied to the prepared case study dataset at pipe level as well as for different strata of pipes according to the primary selected variables for the analysis. Many runs have been made in order to investigate the statistical significance of the primary selected variables and also to check whether the model satisfies the proportionality assumption or not. A PHM has been constructed for each variable or set of variables that found to fulfill the two assumptions of the PHM. Among the three variables that have been examined the number of previous failure (NOPF) was the only variables that found to satisfy the PHM requirements at each run when applied as a continuous variable. A distinct model has been developed for that variable at pipe level and the baseline hazard function has been estimated. Both the hazard function and the baseline hazard function of the model were found to follow the bathtub curve characteristics. However, the variable NOPF was found to be insignificant when included to the model as a discrete variable.

The PHM was found to be applicable at the strata level for three sets of combination of explanatory variables namely: stratified Length adjusted for NOPF, stratified NewEras2 adjusted for NOPF and stratified NOPF adjusted for NewEras2. The three variables were found to be highly statistically significant at all levels of analysis, however the variables Length and NewEras2 were found to violate the proportional hazard assumption for many cases of interaction. It is, thus, recommended to further investigate the time effect of these two variables to accurately capture their multiplicative effect on the hazard of failure for the 100 mm FC pipes in Cape Town.

The baseline hazard functions have been estimated for each Stratified Cox Model (SCM).. The hazard functions and the baseline hazard functions for all models were found to follow the bathtub curve with observed differences in the slope of the curves due to differences in the failure contributing factors. Longer pipes were found to be more prone to failure than shorter once with higher and earlier hazards rates. Similarly, recently installed pipes were found to experience failure more rapidly than older once, indicating to some manufacturing defects.

A summary table for all the PHM models derived in this section to be applied to Cape Town water distribution system is presented below

Derived Proportional Hazard Models for Cape Town's Water Network

The Time Independent Term Summary:

Table 6-7: A summary table for the time independent term for the derived models

	B		SE	Wald	df	Sig	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
PHM at Pipe Level	NOPF	.037	.004	99.450	1	.000	1.038	1.030	1.045
SCM by Length adjusted for NOPF	NOPF	.032	.004	68.698	1	.000	1.033	1.025	1.041
SCM by NewEras2 adjusted for NOPF	NOPF	.042	.004	126.433	1	.000	1.043	1.035	1.051
SCM by NOPF adjusted for NewEras2	New_Era2			537.649	2	.000			
	New_Era2(1)	-4.190	.183	525.928	1	.000	.015	.011	.022
	New_Era2(2)	-1.957	.136	207.707	1	.000	.141	.108	.184

The Baseline Hazard Functions Summary

Table 6-8: A summary for the BHFs for the PHM with NOPF is the covariate

Baseline Hazard Functions, PHM, NOPF

Model period		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
PHM, NOPF (0 < Age <30)	Age (Years)	.004	.001	.299	3.079	.005
	Age (Years) ** 2	.0003	.000	-.609	-2.581	.016
	Age (Years) ** 3	1.922E-005	.000	1.325	8.884	.000
	(Constant)	-.002	.004		-.572	.572
PHM, NOPF (30 < Age <50)	Age (Years)	.071	.004	1.372	16.779	.000
	Age (Years) ** 3	-3.871E-006	.000	-.377	-4.611	.000
	(Constant)	-1.618	.112		-14.415	.000
PHM, NOPF (Age > 50)	Age (Years) ** 2	-.001	.000	-1.388	-4.958	.000
	Age (Years) ** 3	1.617E-005	.000	2.381	8.501	.000
	(Constant)	1.710	.249		6.882	.000

Table 6-9: A summary for the BHFs for the SCM by Length adjusted for NOPF

Baseline Hazard Functions, SCM by L adjusted for NOPF LCatR123

Model period		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
SCM, LCatR1 (1-500 m)	AgeLcatrNOPF	-.005662	.001	-.173	-4.708	.000
	AgeLcatrNOPF ** 2	.000483	.000	1.163	31.633	.000
	(Constant)	.027352	.020		1.378	.172
SCM, LCatR2 (500 – 1000 m)	AgeLCatR2NOPF	-.021	.003	-.458	-6.221	.000
	AgeLCatR2NOPF ** 2	.001	.000	1.429	19.426	.000
	(Constant)	.086	.047		1.825	.075
SCM, LCatR3 (> 1000 m)	AgeLCatR3NOPF	-.011	.004	-.251	-2.485	.019
	AgeLCatR3NOPF ** 2	.001	.000	1.230	12.178	.000
	(Constant)	.040	.058		.693	.494

Table 6-10: A summary for the BHFs for the SCM by NewEras2 adjusted for NOPF

Baseline Hazard Functions, SCM by NewEras adjusted for NOPF

Model period		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
SCM, NewEras1 (1930-1959)	Age1 ** 2	.000	.000	-1.339	-15.438	.000
	Age1 ** 3	9.534E-006	.000	2.325	26.805	.000
	(Constant)	.198	.032		6.110	.000
SCM, NewEras2 (1960-1980)	Age2	-.006	.005	-.117	-1.125	.266
	Age2 ** 2	3.638E-005	.000	.042	.179	.859
	Age2 ** 3	1.703E-005	.000	1.064	7.608	.000
	(Constant)	.044	.042		1.065	.293
SCM, NewEras4 (>1980)	Age3	.06849	.015	1.247	4.556	.000
	Age3 ** 2	-.00624	.001	-3.201	-4.667	.000
	Age3 ** 3	.000223	.000	2.996	6.788	.000
	(Constant)	-.053	.045		-1.174	.253

Table 6-11: A summary for the BHFs for the SCM by NOPF adjusted for NewEras2

Baseline Hazard Functions, SCM by NOPF adjusted for NewEras2

Model period		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
SCM, NOPF (1-4)	AgeNOPF1	.832	.095	.732	8.734	.000
	AgeNOPF1 ** 2	-.043	.003	-2.910	-14.447	.000
	AgeNOPF1 ** 3	.001	.000	3.172	25.053	.000
	(Constant)	-3.047	.823		-3.701	.000
SCM, , NOPF (5-10)	AgeNOPF2	.116	.003	.979	38.052	.000
	(Constant)	.073	.009		7.836	.000
SCM, , NOPF (11 -20)	AgeNOPF3	.126	.004	.980	34.405	.000
	(Constant)	.073	.009		7.743	.000
SCM, , NOPF (> 20)	AgeNOPF4	6.012	2.082	1.939	2.888	.007
	AgeNOPF4 ** 2	-.176	.048	-5.283	-3.662	.001
	AgeNOPF4 ** 3	.002	.000	4.349	5.431	.000
	(Constant)	-65.274	27.701		-2.356	.024

7.3 Recommendations

Based on the results obtained from the study held in this dissertation it is recommended that:

- To maintain the sufficient data that includes and relates all the possible physical and constructional information about their water pipelines which can lead to a much reliable, accurate and informative prediction of the failure patterns in the water distribution system.
- To develop a comprehensive maintenance and rehabilitation strategies for managing the 100 mm FC pipes in Cape Town with the aid of the general failure rate estimates presented in Table 6-2 and the derived Proportional Hazard Models summarized in Tables 6-7, 6-8, 6-9, 6-10, 6-11.
- The effect of pipe lengths on the hazard of failure should be further investigated by considering the time effect of this variable and applying the Extended Cox Model.
- More attention should be paid for
  - 100 mm FC pipes that were installed between 1930 and 1959 as these pipes were found to be at the late stage of wearing out;
  - Longer segments of the 100 mm FC pipes as these pipes tends to have the highest hazard rates and lowest survival probability; and
  - To the quality of manufacturing of new pipes which can affect the reliability of the system.
- To conduct a similar study by including the other FC pipes with different diameters and also different materials within Cape Town's water distribution system to investigate the effect of diameter and material on the hazard of failure of water pipelines in Cape Town.

References

Adey, B., Bernard, O., & Gerard, B. (2003). Risk-Based Replacement Strategies for Deteriorating Reinforced Concrete Pipes. Switzerland: Swiss Federal Institute of Technology.

Alonsoa.C.D., Péreza, R., Lópeza, P. A., & Delgadoa.X. (2008). Influencing Factors for Scheduling Optimal Pipeline Replacement with Budgetary Constraints. *Integrating Sciences and Information Technology for Environmental Assessment and Decision Making* (pp. 1851-1858). Barcelona, Catalonia: Proceedings of the iEMSs Fourth Biennial Meeting. International Congress on Environmental Modelling and Software.

Andreou, S. A. (1986). Predictive Models for Pipe Break Failure and their Implications on Maintenance Planning Strategies for Deteriorating Water Distribution Systems. Phd Thesis. United States of America: Massachusetts Institute of Technology, Dept. of Civil Engineering.

Andreou, S. A., Marks, D. H., & Clark, R. M. (1987). A new Methodology for Modelling Break Failure Patterns in Deteriorating Water distribution Systems: Application. *Advances in Water Resources, vol. 10*, 11-20.

Andreou, S. A., Marks, D. H., & Clark, R. M. (1987). A new Methodology for Modelling Break Failure Patterns in Deteriorating Water distribution Systems: Theory. *Advances in Water Resources, vol. 10*, 2-10.

Basu, A., & Rigdon, S. (2000). *Statistical Methods for the Reliability of Repairable Systems*. New York: John Wiley & Sons, Inc.

Berardi, L. K. (2008). Pipe Deterioration Models for Water Distribution Systems. *Journal of Hydroinformatics (IWA) Publishing,10(2)*, 113-126.

Berardi, L., Giustolisi, O., Kapelan, Z., & Savic, D. A. (2008). Development of Pipe Deterioration Models for Water Distribution Systems using EPR. *Journal of Hydroinformatics, IWA Publishing*, 113-126.

Borovkova, S. (2002). Analysis of survival data. *NAW 5/3 nr*, 302-307.

Bouchan, F., & Goulter, I. (1991). Reliability Improvement in Design of Water Distribution Networks Recognizing Valve Location. *Water Resources Research, vol. 27. No. 12*, 3029-3040.

Boxall, J. B., O'Hagan, A., Pooladsaz, S., & Saul, A. J. (2007). Estimation of Burst Rates in Water Distribution Mains. *Water Management 160*, 73-82.

Breslow, N. (1974). Covariance Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review, vol(43)*, 43—54. In Lee & Wang, 2003.

Burn, L., Eiswirth, M., Desilva, D., & Davis, P. (2001). Condition Monitoring and its Role in Asset Planning. *Pipes Waaga Waaga*, Australia.

Burn, L., Tucker, S., Rahilly, M., Davis, P., Jarrett, R., & Po, M. (2003). Asset Planning for Water Reticulation Systems - The PARMS Model. *Water Science & Technology: Water Supply (3)*, 55-62.

Ciottoni, A. (1983). Computerized Data Management in Determining Causes of Water Main Breaks: The Philadelphia Case Study. *Proceeding of the 1983 International Symposium on Urban Hydrology, Hydraulics and Sediment Control, University of Kentucky, Lexintong, KY.* (pp. 323-329). In: (Goulter, I.C. & Kazemi, A., 1988).

Clark, R. (1982). Water Distribution Systems: A Spatial and Cost Evaluation. *Proceedings of the American Society of Civil Engineers, 108, WR3* (pp. 234-256). Journal of the Water Resources Planning and Management Division.

Cleveland, W. S., & Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association, Vol. 83, No. 403*, 596-610.

COC. (2007, 8 13). *City Allocates Millions to Replace Crumbling Water Pipes*. Retrieved 6 11, 2010, from City of Cape Town Offical Site:
http://www.capetown.gov.za/en/MediaReleases/Pages/CityAllocatesMillionsToReplaceCrumblingWaterPipes.aspx

COC. (2007a, October). Reticulation 8 Districts and Major Roads. Cape Town: City of Cape Town.

COC. (2008). Water Services Development Plan,WSDP. Cape Town: City of Cape Town.

COC. (2009, 5 26). *City Upgrading Water Infrastructure*. Retrieved 6 2, 2010, from City of Cape Town Official Site:
http://www.capetown.gov.za/en/Pages/Cityupgradingwaterinfrastructure.aspx

COC. (2011, April 14). *Cape Town's Future Development Plans near Completion.* Retrieved December 6, 2011, from City of Cape Town Official Site: http://www.capetown.gov.za/en/Pages/CTsfuturedevelopmentnearcompletion.aspx

COC. (2011). Water Services – Reticulation, Aging and Relay of Water & Sewage Reticulation. Cape Town: City of Cape Town.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* Second Edition. Chapman & Hall/CRC.

Collicot.K.G. (2004). Best Practices for Water Distribution Systems. *New Developments in Material and Technology Simposium*. Canada: R.V. Anderson Associates Limited and National Research Council.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2.*, 187-220.

Cox, D., & Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

Creig-Smith, S. (2001). To Refurbish or Replace Steel Water Pipelines, that is the Question. *Engineering Failure Analysis (8)*, 107-112.

Czyz, J. A., & Pettigrew, C. (2000). Multi-Pipeline Geographical Information System Based on High Accuracy Inertial Surveys. *Proceedings of International Pipeline Conference, IPC*, (pp. 00—138). Mexico City.

Davis, P., Burn, S., Moglia, M., & Gould, S. (2007). A Physical Probabilistic Model to Predict Failure Rates in Buried PVC Pipelines. *Reliability Engineering and System Safety, 92*, 1258–1266.

DelMistro. (2011). Personal Communication.

DeSilva.D, Davis.P, Burn.L.S, Ferguson.P, Massie.D, Cull.J, et al. (2003). *Condition Assessment of Cast Iron and Asbestos Cement Pipes by In-Pipe Probes and Selective Sampling for Estimation of Remaining Life*. Australia: CSIRO.

Eiswirth, M., Heske, C., Burn, L., & DeSilva, D. (2001). New Methods of Water Pipeline Assessment. *World Water Congress*. Berlin, Germany: IWA (2).

Engelhardt, M., Skipworth, P., Savic, D., & A.J. Saul b, G. W. (2000). Rehabilitation Strategies for Water Distribution Networks: a Literature Review with a UK Perspective. *Urban Water (2)*, 153-170.

Geehman, C. (1999). Prioritising Water Main Renewals. Wagga Wagga: Pipes '99.

Goulter.I.C, & Kazemi.A. (1988). Spatial and Temporal Grouping of Water Main Pipe Breakage in Winnipeg. *Candian Journal of Civil Engineering, 15*, 91-97.

Graybill, F., & Iyer, H. (1994). *Regression Analysis: Concepts and Applications*. Belmont, California: Duxbury Press.

Gustafson, J.-M., & Clancy, D. V. (1999). Modeling the Occurrence of Breaks in Cast Iron Water Mains Using Methods of Survival Analysis. *Annual Conference Proceedings June 20-24*. Chicago: American Water Works Association.

Hadzilacos, T., Kalles, D., Preston, N., Melbourne, P., Camarinopoulo, L., Eimermache, M., et al. (2000). UTILNETS: A Water Mains Rehabilitation Decision Support System. *Computers, Environment and Urban Systems (24)*, 215-232.

Herz, R. K. (1998). Exploring Rehabilitation Needs and Strategies for Water Distribution Networks. *J Water SRT - Aqua (47)*, 275-283.

Herz, R., & Kroop, I. (2002). CARE-W Computer Aided Rehabilitation of Water Networks. Decision Support Tools for Sustainable Water Network Management. Germany: Research and Technological Development Project of European Community.

Hooper, M. R. (2010). Statistical Modelling of Water Pipeline Failure. Msc Thesis. Cape Town: University of Cape Town.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* United States: John Wiley & Sons, Inc.

Hu, Y., & Hubble, D. (2007). Factors Contributing to the Failure of Asbestos Cement Water Mains. *National Research Council*, 608-621.

Ingenium. (2006). *International Infrastructure Management Manual.* New Zeland: Association of Local Government Engineering N.Z. Inc (INGENIUM), National Asset Management Steering (NAMS) Group.

Jamison.j.M, Shawn.M.O, & Patrick.E.W. (2001, December 15). An Examination of Methods for Condition Rating of Sewer Pipelines. Minnesota, United States.

Jarvis, B. (1998). Asbestos-cement-pipe corrosion, interim report. Australia: Customer Service Division, Water Corporation,Westren Australia. In: Y.Hu & Hubble, 2007.

Jeffrey, L. A. (1985). Predicting Urban Water Distribution Maintenance Strategies: A Case Study of New Haven, Connecticut. Msc Thesis. Massachusetts Institute of Technology, Dept. of Civil Engineering.

Jenkins, S. P. (2005, July 18). *Survival Analysis.* Retrieved 1 13, 2012, from http://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* Second Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.

Kettler, A., & Goulter, I. (1985). An Analysis of Pipe Breakage in Urban Water Distribution Networks. *Canadian Journal of Civil Engineering*, 286-293.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censord and Truncated Data, Second Edition.* New York: Springer.

Kleinbaum, D. G., & Klein, M. (2005). *Survival Analysis A Self-Learning Text. Second Edition.* United States: Springer Science Inc.

Kleiner, Y. (1997). *Rehabilitation Planning of Water Distribution Networks : The Component and the System Perspective.* Canada: National Research Council.

Kleiner, Y. (1997). Water Distribution Network Rehabilitation: Selection and Scheduling of Pipe Rehabilitation Alternatives. Phd. Thesis. Canda: University of Toronto.

Kleiner, Y., & Rajani, B. (2000). Considering Time-dependent Factors in the Statistical Prediction of Water Main Breaks. *AWWA Infrastructure Conference Proceedings, March 12-15, 2000,* (pp. 1-12). Baltimore, Maryland: Institute for Research in Construction, National Research Council Canada.

Kleiner, Y., & Rajani, B. (2001). Comprehensive Review of Structural Deterioration of Water Mains: Statistical Models. *Urban Water Journal (3)*, 131-150.

Kleiner, Y., & Rajani, B. (2010). I-WARP: Individual Water Main Renewal Planner. *Drinking Water Engineering and Science Discussions (3)*, 25–41.

Kleiner, Y., Adams, B. J., & Rogers, J. (1998a). Long-Term Planning Methodology for Water Distribution System Rehabilitation. *Water Resources Research, 34(8)*, 2039–2051.

Kleiner, Y., Adams, B., & Rogers, J. (1998b). Selection and Scheduling of Rehabilitation Alternatives for Water Distribution Systems. *Water Resources Research (34)*, 2053-2061.

Kleiner, Y., Rajani, B., & Wang, S. (2007). *Consideration of Static and Dynamic Effects to Plan Water Main Renewal.* Ottawa, Ontario, Canada: National Research Council Canada (NRC).

Kleiner, Y., Sadiq, R., & Rajani, B. (2006). Modelling the Deterioration of Buried Infrastructure as a Fuzzy Markov Process. *Institute for Research in Construction, National Research Council Canada (NRC)*, Ottawa, Canada.

Kleiner, Y., Tesfamariam, S., Rajani, B., & Abdel-Akher, A. (2007). Distribution-Water Mains Renewal Planner. Ottawa, Canada: National Research Council.

Kumar, D., & Klefsjö, B. (1994). Proportional Hazards Model - A Review. *Reliability Engineering & System Safety,vol 44*, 177-188. In (Røstum, 2000).

Lee, E. T., & Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis, Third Edition.* Hoboken, New Jersey: John Wiley & Sons, Inc.

LeGat, Y., & Eisenbeis, P. (2000). Using Maintenance Records to Forecast Failures in Water Networks. *Urban Water (2)*, 173-181.

Machin, D., Cheung, Y. B., & Parmar, M. K. (2006). *Survival Analysis: A Practical Approach.* Chichester, United Kingdom: John Wiley & Sons Ltd.

Makar.J, & Kleiner.Y. (2000). Maintaining Water Pipeline Integrity. *National Research Council.*

McGrath, T. J., & Mruk, S. A. (2002). Thermoplastic Piping. In P. Mohinder L. Nayyar, *Piping Handbook.* United States: McGraw-Hill.

Misiunas, D. (2005). Failure Monitoring and Asset Condition Assessment in Water Supply Systems. Doctoral Dissertation in Industrial Automation. Lund, Southern Sewden: Department of Industrial Electrical Engineering and Automation, Lund University.

Moglia, M., Burn, S., & Meddings, S. (2006). Decision Support System for Water Pipeline Renewal Prioritisation. *ITcon Vol. 11* , pg. 237.

Morris, R. (1975). The Distribution System Manual of Water Utility Operations. Tex. Austin, United States: Tex.Water Utilities Bd. In: Walski and Pelliccia 1982.

NAMS, T. N. (2002). *International Infrastructure Management Manual (IIMM).* New Zeland: Association of Local Government Engineering NZ Inc (INGENIUM).

Nathabandu, T. K., & Rosso, R. (2008). *Applied Statistics for Civil and Environmental Engineers.* Second Edition. United Kingdom: Blackwell Publishing Ltd.

Nelson, W. (1988). Graphical Analysis of System Repair Data. *Journal of Quality Technology, 20,* 24-35.

Nelson, W. (1995). Confidence Limits for Recurrence Data: Applied to Cost or Number of Product Repairs . *Technometrics, Vol. 37, No. 2,* 147-157.

Nelson, W. B. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications* . Schenectady, New York : The American Statistical Association (ASA) and the Society for Industrial and Applied Mathematics (SIAM).

Nesbar, B. (1983). *Asbestos/Cement Pipe Corrosion: Part 2- Review of Recent Work on the Causes of Pipe Degradation and on Possible Improvements.* Ottawa, Ontario. : CANMET Report 83-17E, Canada Centre for Mineral and Energy Technology, Energy, Mines and Resources. In: Y.Hu & Hubble, 2007.

NRC. (2003). Deterioration and Inspection of Water Distribution Systems. Canada: Federation of Canadian Municipalities and National Research Council.

NRC. (2006). *Drinking Water Distribution Systems Assessment and Reducing Risks.* Washington: National Research Council of the National Academies.

O' Day, D. (1988). Water Main Condition Monitering-A Case Study. *Pipeline Infrastructure Proceeding* (pp. 274-286). In: (DeSilva, D. et al., 2003).

O'Day, D. (1982). Organizing and Analyzing Leak and Break Data for Making Main Replacement Decisions. *Journal of the American Water Works Association, 74,* 589-596.

O'Day.D.K, e. a. (1980). Aging Urban Water Systems:A Computerized Case Study. *Public Work, 8*(3), 61-111. In: (Goulter. I. C & Kazemi.A, 1988).

Park, S. (2004). Identifying the Hazard Characteristics of Pipes In Water Distribution Systems by using the Proportional Hazard Model: 2 Application. *KSCE Journal of Civil Engineering, 8(6)*, 669-677.

Park, S., Jun, H., Agbenowosi, N., Kim, B. J., & Lim, K. (2010). The Proportional Hazards Modeling of Water Main Failure Data Incorporating the Time-dependent Effects of Covariates. *Water Resources Management*, DOI 10.1007/s11269-010-9684-y.

Park, S., Jun, H., Kim, B. J., & Im, G. C. (2008). Modeling of Water Main Failure Rates Using the Log-linear ROCOF and the Power Law Process. *Water Resources Management, 22*, 1311–1324.

Pelletier, G., Mailhot, A., & Villeneuve, J.-P. (2003). Modeling Water Pipe Breaks—Three Case Studies. *Journal of Water Resources Planning and Management © ASCE*, 115-123.

Peter D. Rogers, P., & and Neil S. Grigg, M. (2009). Failure Assessment Modeling to Prioritize Water Pipe Renewal: Two Case Studies. *Journal of Infrastructure Systems© ASCE*, 162-171.

Poulton, M., Gat, Y. L., & Brémond, B. (2007). The Impact of Pipe Segment Length on Break Predictions in Water Distribution Systems. *2nd Leading Edge Conference on Strategic Asset Management.* Lisbon, Portugal: International Water Association.

Priddy, D. (2008). Why Do PVC and CPVC Pipes Fail? Midland, United States: Plastic Failure Labs.

Rajani, B., & Kleiner, Y. (2001). Comprehensive Review of Structural Deterioration of Water Mains: Physically Based Models. *Urban Water (3)*, 151-164.

Rajani, B., & Tesfamariam, S. (2007). Estimating Time to Failure of Cast-Iron Water Mains. *Proceedings of the ICE - Water Management. 160*, pp. pages 83 –88. Institute for Research in Construction, National Research Council Canada.

Rajani, Zhan, & Kuraoka. (1996). Pipe Soil Interaction Analysis of Jointed Water Mains. *Institute for Research in Construction, National Research Council of Canada, NRC*, 393-404.

Rajani.B, & Kleiner.Y. (2003). *Protection of Ductile Iron Water Mains against External Corrosion: Review of Methods and Case Histories.* Canada: Institute for Research in Construction, National Research Council Canada.

Røstum, J. (2000). Statistical Modelling of Pipe Failures in Water Networks. Phd Thesis. Trondheim, Norway: Faculty of Civil Engineering,the Norwegian University of Science and Technology.

Saegrov.S, Baptista, J. M., Conroy, P., Herz, R., LeGaufre, P., Moss, G., et al. (1999). Rehabilitation of Water Networks Survey of Research Needs and On-going Efforts. *Urban Water, volume (1)*, 15-22.

Shahata, K., & Zayed, T. (2008). Simulation as a Tool for Lifecycle Cost Analysis. *Winter Simulation Conference, 2008*, (pp. 2497-2503). Canada.

Shamir, U., & Howard, C. D. (1979). An Analytic Approach to Scheduling Pipe Replacement. *Amirican Water Works Association*, 248-258.

Shoukri, M., & Pause, C. (1999). *Statistical Methods for Health Sciences*. Second Edition. United States: CRC Press LLC.

Skipworth, P. E. (2002). *Whole Life Costing for Water Distribution Network Management*. London: Thomas Telford.

Soong, T. (2004). *Fundamentals of Probabililty and Statistics for Engineers*. United Kingdom: John Wiley & Sons Ltd.

Srikanth.S, e. a. (2005). Corrosion in a Buried Pressurised Water Pipeline. *Engineering Failure Analysis (12)*, 634–651.

Stacha, J. (1978). Criteria for Pipeline Replacement. *American Water Work Association, AWWA*, 256-258.

Stone, S., Dzuray, E. J., Meisegeier, D., Dahlborg, A., & Erickson, M. (2007). Decision-Support Tools for Predicting the Performance of Water Distribution and Wastewater Collection Systems. Cincinnati: U.S. Environmental Protection Agency.

Thornton, J., Sturm, R., & George Kunkel, P. (2008). *Water Loss Control*. Second Edition. United States: Mc Grow Hill.

Tomboulian, P., Schweitzer, L., Mullin, K., J.Wilson, & D.Khiari. (n.d.). Material Used in Water Distribution Systems: Contribution to Tastes- and- Odour. *Water Science and Technology, 49*, 219-226.

Uni-Bell. (2003). *PVC Pressure Pipe: The Solution For Water Systems* . Retrieved 6 9, 2010, from Uni-Bell PVC Pipe Association:
http://www.plastics.org.nz/pipa/attachments/docs/pvc-advantages-uni-bell.pdf

Walski, T., & Pelliccia, A. (1982). Economic Analysis of Water Mains Breaks. *American Water Works Association (AWWA), 74*, 140-147.

Wasson, C. (2006). *System Analysis, Design, and Development: Concepts, Principles, and Practices.* Canada: John Wiley and Sons Inc.

Watson, T., Christian, C., Mason, A., & Smith, M. (2001). Maintenance of Water Distribution Systems. New Zealand: The University of Auckland.

Wengström, T. (1993). Drinking Water Pipe Breakage Records: A Tool for Evaluating Pipe and System Reliability. *Institutionen för vattenförsörjnings- och avloppsteknik,Chalmers tekniska högskola.* , In: (Røstum, 2000).

Wilkins, D. J. (2002a). *The Bathtub Curve and Product Failure Behaviour. Part One - The Bathtub Curve, Infant Mortality and Burn-in.* Retrieved 7 14, 2010, from Wiebull.Com. Reliability Engineering Resources:
http://www.weibull.com/hotwire/issue21/hottopics21.htm

Wilkins, D. J. (2002b). *The Bathtub Curve and Product Failure Behaviour. Part Two - Normal Life and Wear-Out.* Retrieved 7 14, 2010, from weibull.com. Reliability Engineering Resource: http://www.weibull.com/hotwire/issue22/hottopics22.htm

Wood, A., & Lence, B. J. (2009). Using Water Main Break Data to Improve Asset Management for Small and Medium Utilities: District of Maple Ridge, B.C. *Journal of Infrastructure Systems , ASCE*, 111-119.

Workman, L. (2009). *PVC Failure Analysis.* Retrieved 6 2010, from Expert4PVC Consulting: http://expert4pvc.com/Documents/Failure_Analysis_101.pdf

WSAA. (2003). *Common Failure Modes in Pressurized Pipeline Systems.* Retrieved 6 23, 2010, from Water Services Association of Australia:
https://www.wsaa.asn.au/Publications/NationalCodes/Documents/Failure%20Mode%20Table%20Compendia%2002.pdf

Yaminighaeshi, H. (2003). Probability of Failure Analysis and Condition Assessment of Cast Iron Pipes due to Internal and External Corrosion in Water Distribution Systems. Phd thesis. Vancouver: University of British Columbia.

Zamojc.M.D. (2005). Material Specification of Large Diameter Water Mains. Canada: Regional Municipality of Peel.

Appendix

PHM dataset

Age	Installation Era	Censorship	NOPF	L	LCatsMod	LnL	SqrtL	LCatsR	NewNOPF	NewEra2
0	9	0	1	399	2	5.99	19.97	1	1	3
0	11	0	1	405	3	6	20.12	1	1	3
0	10	0	1	337	2	5.82	18.36	1	1	3
0	9	1	2	76	1	4.33	8.72	1	1	3
0	9	0	1	429	3	6.06	20.71	1	1	3
0	9	1	2	6	1	1.79	2.45	1	1	3
0	9	1	2	166	1	5.11	12.88	1	1	3
0	10	1	5	300	2	5.7	17.32	1	2	3
1	9	0	1	387	2	5.96	19.67	1	1	3
1	9	0	1	167	1	5.12	12.92	1	1	3
1	10	1	4	158	1	5.06	12.57	1	1	3
1	9	0	2	166	1	5.11	12.88	1	1	3
1	9	1	12	240	2	5.48	15.49	1	3	3
1	9	0	1	161	1	5.08	12.69	1	1	3
2	9	0	1	338	2	5.82	18.38	1	1	3
2	9	0	1	88	1	4.48	9.38	1	1	3
2	9	0	1	601	3	6.4	24.52	2	1	3
2	10	1	4	158	1	5.06	12.57	1	1	3
...	...	...	...	...	...	...	...	...	...	...

PHM at Pipe Level

PHM, NOPF is the Covariate

Survival Table

Time	Baseline Cum Hazard	At mean of covariates		
		Survival	SE	Cum Hazard
0	.002	.998	.001	.002
1	.003	.997	.001	.003
2	.003	.996	.002	.004
3	.004	.995	.002	.005
4	.007	.991	.002	.009
5	.010	.987	.003	.013
6	.011	.986	.003	.014
7	.012	.985	.003	.016
8	.013	.983	.003	.017
9	.016	.980	.003	.021
10	.023	.971	.004	.030
11	.026	.967	.004	.034
12	.027	.966	.004	.035
13	.031	.960	.005	.041
14	.040	.949	.005	.052
15	.049	.937	.006	.065
16	.058	.926	.006	.077
17	.068	.914	.007	.090
18	.079	.901	.007	.104
19	.087	.892	.007	.114
20	.097	.881	.008	.127
21	.111	.864	.008	.146
22	.119	.855	.008	.156
23	.128	.845	.009	.168
24	.146	.825	.009	.192
25	.165	.805	.010	.217
26	.191	.777	.010	.252
27	.211	.758	.010	.277
28	.234	.734	.011	.309
29	.272	.699	.011	.358
30	.307	.668	.012	.404
31	.342	.637	.012	.450
32	.392	.597	.012	.517

33	.428	.569	.012	.564
34	.464	.543	.012	.611
35	.503	.515	.012	.663
36	.549	.485	.013	.723
37	.587	.462	.013	.773
38	.630	.436	.013	.830
39	.672	.413	.013	.885
40	.713	.391	.013	.939
41	.749	.373	.013	.986
42	.805	.346	.012	1.060
43	.837	.332	.012	1.102
44	.883	.313	.012	1.162
45	.916	.299	.012	1.206
46	.945	.288	.012	1.244
47	.986	.273	.012	1.299
48	1.032	.257	.012	1.358
49	1.035	.256	.012	1.363
50	1.055	.249	.012	1.389
51	1.083	.240	.012	1.426
52	1.112	.231	.012	1.464
53	1.162	.217	.012	1.530
54	1.183	.211	.012	1.557
55	1.231	.198	.012	1.621
56	1.265	.189	.012	1.665
57	1.317	.176	.012	1.735
58	1.357	.167	.011	1.787
59	1.433	.152	.011	1.887
60	1.440	.150	.011	1.896
61	1.500	.139	.011	1.975
62	1.525	.134	.011	2.008
63	1.578	.125	.011	2.078
64	1.616	.119	.011	2.128
65	1.690	.108	.010	2.225
66	1.815	.092	.010	2.390
67	1.872	.085	.010	2.465
68	1.942	.078	.009	2.557
69	1.961	.076	.009	2.582
70	2.024	.070	.009	2.665
71	2.126	.061	.009	2.800

72	2.158	.058	.009	2.842
73	2.349	.045	.008	3.093
74	2.457	.039	.008	3.235
76	2.527	.036	.008	3.327
77	2.701	.029	.008	3.556

SCM by length adjusted for NOPF

Survival Table

Time		Baseline Cum Hazard	At mean of covariates		
			Survival	SE	Cum Hazard
LengthCategoriesRecent=0-500	0	.003	.997	.002	.003
	1	.004	.995	.002	.005
	2	.005	.993	.002	.007
	3	.006	.993	.002	.007
	4	.009	.988	.003	.012
	5	.012	.984	.004	.016
	6	.014	.983	.004	.018
	7	.015	.981	.004	.019
	8	.017	.979	.004	.021
	9	.018	.978	.004	.023
	10	.025	.968	.005	.032
	11	.027	.966	.005	.035
	12	.028	.965	.005	.036
	13	.033	.959	.006	.042
	14	.040	.950	.006	.051
	15	.053	.934	.007	.068
	16	.064	.922	.008	.081
	17	.070	.915	.008	.089
	18	.078	.906	.008	.099
	19	.084	.899	.009	.107
	20	.095	.887	.009	.120
	21	.107	.872	.010	.136
	22	.114	.865	.010	.145
	23	.122	.856	.010	.155

	24	.139	.838	.011	.176
	25	.152	.824	.011	.194
	26	.175	.801	.012	.222
	27	.193	.782	.012	.245
	28	.217	.759	.013	.276
	29	.255	.724	.013	.324
	30	.294	.688	.014	.374
	31	.329	.658	.014	.418
	32	.374	.621	.014	.476
	33	.408	.595	.015	.519
	34	.437	.574	.015	.556
	35	.461	.556	.015	.586
	36	.503	.528	.015	.639
	37	.545	.500	.015	.693
	38	.574	.482	.015	.730
	39	.608	.462	.015	.772
	40	.645	.441	.015	.820
	41	.674	.424	.015	.857
	42	.724	.398	.015	.920
	43	.749	.386	.015	.951
	44	.783	.370	.015	.995
	45	.808	.358	.015	1.027
	46	.841	.344	.015	1.068
	47	.875	.329	.015	1.112
	48	.920	.311	.015	1.169
	49	.923	.309	.015	1.173
	50	.942	.302	.015	1.197
	51	.973	.290	.015	1.236
	52	.997	.282	.015	1.267
	53	1.053	.262	.015	1.338
	54	1.076	.255	.015	1.367
	55	1.130	.238	.015	1.436
	56	1.167	.227	.014	1.483
	57	1.214	.214	.014	1.543

	58	1.252	.204	.014	1.591
	59	1.320	.187	.014	1.678
	60	1.328	.185	.014	1.687
	61	1.393	.170	.014	1.770
	62	1.419	.165	.014	1.804
	63	1.476	.153	.013	1.875
	64	1.515	.146	.013	1.926
	65	1.594	.132	.013	2.025
	66	1.726	.112	.012	2.193
	67	1.787	.103	.012	2.271
	68	1.861	.094	.012	2.365
	69	1.881	.092	.012	2.391
	70	1.950	.084	.011	2.478
	71	2.061	.073	.011	2.619
	72	2.095	.070	.011	2.662
	73	2.298	.054	.010	2.920
	74	2.411	.047	.010	3.065
	76	2.484	.043	.010	3.157
	77	2.672	.034	.010	3.395
LengthCategoriesRecent =500-1000	4	.002	.998	.002	.002
	5	.003	.996	.003	.004
	6	.005	.993	.004	.007
	8	.007	.991	.004	.009
	9	.012	.985	.006	.015
	10	.017	.978	.007	.022
	11	.021	.974	.007	.026
	12	.023	.972	.008	.029
	14	.028	.965	.009	.035
	16	.033	.959	.009	.042
	17	.044	.945	.011	.056
	18	.061	.926	.012	.077
	19	.076	.908	.014	.096
	20	.087	.895	.014	.111
	21	.109	.870	.016	.139
	22	.124	.854	.017	.157
	23	.137	.841	.017	.174

	24	.165	.811	.019	.210
	25	.200	.776	.020	.254
	26	.232	.745	.021	.294
	27	.252	.726	.022	.320
	28	.273	.707	.022	.347
	29	.318	.668	.023	.404
	30	.344	.646	.023	.438
	31	.391	.609	.024	.497
	32	.459	.558	.025	.583
	33	.505	.526	.025	.642
	34	.566	.487	.025	.719
	35	.651	.437	.025	.827
	36	.696	.413	.025	.884
	37	.732	.394	.025	.931
	38	.822	.352	.025	1.044
	39	.919	.311	.024	1.168
	40	.986	.286	.024	1.253
	41	1.065	.258	.024	1.354
	42	1.146	.233	.023	1.457
	43	1.240	.207	.023	1.576
	44	1.363	.177	.022	1.733
	45	1.470	.154	.022	1.869
	47	1.639	.125	.021	2.083
	48	1.738	.110	.021	2.209
	50	1.803	.101	.021	2.291
	52	1.982	.081	.020	2.519
	58	2.122	.067	.020	2.697
	59	2.594	.037	.016	3.296
LengthCategoriesRecent =>1000	4	.006	.992	.008	.008
	5	.012	.984	.011	.016
	9	.019	.977	.013	.024
	10	.031	.961	.017	.040
	11	.044	.945	.020	.056
	13	.064	.922	.024	.082
	14	.106	.874	.029	.135
	15	.121	.857	.031	.154
	16	.144	.833	.033	.183

	17	.193	.782	.036	.245
	18	.219	.757	.038	.278
	21	.228	.748	.039	.290
	23	.238	.739	.039	.302
	25	.266	.713	.041	.338
	26	.338	.651	.043	.430
	27	.383	.614	.044	.487
	28	.432	.578	.045	.549
	29	.470	.550	.045	.598
	30	.511	.522	.046	.649
	31	.525	.513	.046	.667
	32	.601	.466	.046	.764
	33	.652	.436	.046	.829
	34	.708	.407	.046	.900
	35	.815	.355	.045	1.036
	36	.976	.289	.042	1.241
	38	1.116	.242	.040	1.418
	40	1.157	.230	.040	1.471
	41	1.200	.218	.040	1.525
	42	1.376	.174	.037	1.748
	44	1.571	.136	.036	1.997
	45	1.709	.114	.035	2.172
	57	2.328	.052	.026	2.959

SCM by NewEras2 adjusted for Length

Survival Table

Time		Baseline Cum Hazard	At mean of covariates		
			Survival	SE	Cum Hazard
	36	.010	.988	.007	.012
	37	.017	.981	.008	.019
New_Era2=1930-1959	38	.031	.965	.011	.035
	39	.037	.958	.012	.043
	40	.048	.946	.014	.055

	41	.059	.935	.015	.067
	42	.088	.904	.018	.101
	43	.091	.900	.018	.105
	44	.114	.877	.020	.131
	45	.130	.861	.021	.149
	46	.146	.846	.022	.167
	47	.170	.822	.024	.196
	48	.195	.799	.025	.224
	50	.217	.779	.026	.250
	51	.249	.751	.027	.286
	52	.277	.727	.028	.318
	53	.341	.675	.029	.392
	54	.367	.655	.029	.423
	55	.423	.615	.030	.487
	56	.465	.586	.031	.535
	57	.523	.548	.031	.602
	58	.572	.518	.031	.658
	59	.664	.466	.031	.764
	60	.672	.462	.031	.773
	61	.745	.424	.031	.857
	62	.775	.410	.031	.892
	63	.841	.380	.031	.968
	64	.889	.360	.031	1.022
	65	.982	.323	.031	1.130
	66	1.139	.270	.029	1.311
	67	1.212	.248	.029	1.395
	68	1.303	.223	.028	1.499
	69	1.328	.217	.028	1.528
	70	1.413	.197	.028	1.626
	71	1.554	.167	.027	1.788
	72	1.601	.158	.027	1.842
	73	1.877	.115	.024	2.160
	74	2.025	.097	.023	2.329
	76	2.116	.088	.023	2.434

	77	2.390	.064	.024	2.749
	4	.001	.999	.001	.001
	5	.002	.998	.001	.002
	6	.003	.997	.001	.003
	8	.004	.996	.002	.004
	9	.006	.993	.002	.007
	10	.012	.986	.003	.014
	11	.016	.982	.004	.018
	12	.017	.981	.004	.019
	13	.023	.974	.004	.026
	14	.033	.963	.005	.038
	15	.042	.952	.006	.049
	16	.053	.941	.006	.061
	17	.066	.926	.007	.076
	18	.082	.910	.008	.094
	19	.093	.898	.008	.108
	20	.106	.885	.009	.122
	21	.125	.866	.009	.144
	22	.135	.856	.010	.155
	23	.146	.845	.010	.169
New_Era2=1960-1989	24	.168	.824	.010	.193
	25	.194	.800	.011	.224
	26	.231	.766	.012	.266
	27	.259	.742	.012	.298
	28	.295	.712	.012	.339
	29	.352	.667	.013	.405
	30	.405	.627	.013	.466
	31	.460	.589	.014	.529
	32	.541	.537	.014	.622
	33	.600	.501	.014	.690
	34	.662	.467	.014	.761
	35	.730	.432	.014	.840
	36	.808	.395	.014	.929
	37	.875	.365	.014	1.007
	38	.953	.334	.014	1.097
	39	1.037	.303	.014	1.193
	40	1.120	.276	.013	1.288
	41	1.196	.253	.013	1.376
	42	1.309	.222	.013	1.506

	43	1.399	.200	.013	1.609
	44	1.515	.175	.012	1.743
	45	1.614	.156	.012	1.857
	46	1.710	.140	.012	1.968
	47	1.877	.115	.012	2.159
	48	2.128	.086	.012	2.448
	49	2.171	.082	.012	2.498
	50	2.221	.078	.012	2.555
	51	2.305	.071	.013	2.652
	52	2.529	.055	.013	2.909
	55	2.728	.043	.014	3.139
	57	3.365	.021	.013	3.872
New_Era2=1990-2010	0	.027	.969	.015	.031
	1	.041	.953	.018	.048
	2	.056	.937	.021	.065
	3	.064	.929	.023	.074
	4	.105	.887	.028	.120
	5	.156	.836	.033	.179
	6	.174	.819	.034	.200
	7	.192	.802	.036	.221
	8	.202	.793	.036	.232
	9	.221	.775	.038	.255
	10	.295	.712	.041	.340
	11	.307	.702	.042	.353
	12	.319	.693	.042	.367
	13	.332	.683	.043	.382
	14	.374	.650	.044	.430
	15	.483	.574	.047	.556
	16	.554	.529	.048	.638
	17	.596	.504	.049	.686
	20	.679	.458	.050	.781
	21	.775	.410	.052	.892
	22	.852	.375	.052	.981
	23	.941	.339	.053	1.082
	24	1.301	.224	.047	1.497
	25	1.481	.182	.045	1.704

26	1.845	.120	.038	2.123
27	2.096	.090	.036	2.412

SCM NewEras2 adjusted for NOPF

Survival Table

Time		Baseline Cum Hazard	At mean of covariates		
			Survival	SE	Cum Hazard
New_Era2=1930-1959	36	.008	.989	.006	.011
	37	.013	.982	.008	.018
	38	.024	.968	.011	.033
	39	.030	.960	.012	.040
	40	.038	.949	.013	.052
	41	.046	.939	.014	.063
	42	.069	.910	.017	.095
	43	.072	.906	.017	.099
	44	.090	.884	.019	.123
	45	.102	.869	.020	.140
	46	.115	.855	.021	.157
	47	.134	.833	.022	.183
	48	.153	.811	.023	.210
	50	.170	.792	.024	.233
	51	.194	.766	.025	.266
	52	.216	.744	.026	.295
	53	.265	.696	.027	.363
	54	.285	.677	.028	.390
	55	.328	.638	.029	.449
	56	.361	.611	.029	.493
	57	.406	.574	.030	.555
	58	.444	.545	.030	.607
	59	.517	.493	.030	.707
	60	.523	.489	.030	.716
	61	.581	.452	.031	.795
	62	.605	.437	.031	.827

	63	.655	.408	.031	.896
	64	.692	.388	.031	.946
	65	.762	.353	.031	1.042
	66	.881	.300	.030	1.205
	67	.936	.278	.029	1.281
	68	1.002	.254	.029	1.370
	69	1.020	.248	.029	1.395
	70	1.080	.228	.029	1.477
	71	1.176	.200	.028	1.609
	72	1.206	.192	.028	1.650
	73	1.385	.151	.026	1.894
	74	1.486	.131	.025	2.033
	76	1.552	.120	.025	2.123
	77	1.713	.096	.027	2.342
New_Era2=1960-1989	4	.001	.999	.001	.001
	5	.002	.998	.001	.002
	6	.002	.997	.001	.003
	8	.003	.996	.002	.004
	9	.005	.993	.002	.007
	10	.010	.987	.003	.013
	11	.013	.982	.003	.018
	12	.013	.982	.004	.018
	13	.018	.975	.004	.025
	14	.027	.964	.005	.037
	15	.034	.954	.006	.047
	16	.043	.943	.006	.059
	17	.054	.929	.007	.073
	18	.066	.913	.007	.091
	19	.076	.902	.008	.103
	20	.086	.889	.008	.117
	21	.101	.871	.009	.138
	22	.109	.861	.009	.149
	23	.119	.850	.010	.162
	24	.136	.830	.010	.186
	25	.158	.806	.011	.216
	26	.188	.773	.011	.257
	27	.211	.749	.012	.289

	28	.240	.720	.012	.329
	29	.287	.675	.013	.393
	30	.332	.635	.013	.454
	31	.377	.597	.014	.516
	32	.445	.544	.014	.608
	33	.494	.509	.014	.676
	34	.547	.473	.014	.748
	35	.606	.437	.014	.828
	36	.673	.399	.014	.920
	37	.731	.368	.014	1.000
	38	.799	.335	.014	1.092
	39	.872	.304	.014	1.192
	40	.944	.275	.013	1.292
	41	1.012	.250	.013	1.384
	42	1.114	.218	.013	1.524
	43	1.196	.195	.013	1.636
	44	1.303	.168	.012	1.782
	45	1.393	.149	.012	1.905
	46	1.481	.132	.012	2.025
	47	1.631	.107	.012	2.231
	48	1.859	.079	.011	2.543
	49	1.898	.075	.011	2.596
	50	1.944	.070	.011	2.659
	51	2.025	.063	.012	2.770
	52	2.242	.047	.013	3.066
	55	2.436	.036	.013	3.332
	57	3.074	.015	.011	4.204
New_Era2=1990-2010	0	.024	.968	.015	.032
	1	.036	.952	.019	.049
	2	.049	.935	.022	.067
	3	.056	.927	.023	.076
	4	.090	.884	.029	.124
	5	.134	.833	.034	.183
	6	.149	.815	.035	.204
	7	.165	.798	.036	.226
	8	.173	.789	.037	.237
	9	.190	.771	.038	.260

	10	.253	.708	.041	.346
	11	.263	.698	.042	.359
	12	.273	.689	.042	.373
	13	.283	.679	.043	.388
	14	.318	.647	.044	.435
	15	.409	.572	.047	.559
	16	.469	.527	.048	.641
	17	.504	.502	.048	.689
	20	.570	.459	.050	.780
	21	.648	.412	.051	.887
	22	.710	.379	.051	.971
	23	.780	.344	.052	1.067
	24	1.073	.230	.046	1.468
	25	1.221	.188	.045	1.670
	26	1.520	.125	.038	2.079
	27	1.724	.095	.037	2.358

Survival Table (SCM by NOPF adjusted for NewEras2)

Time		Baseline Cum Hazard	At mean of covariates		
			Survival	SE	Cum Hazard
New NOPF=1-4	0	.019	.998	.001	.002
	1	.026	.997	.002	.003
	2	.040	.995	.002	.005
	3	.047	.995	.002	.005
	4	.069	.992	.003	.008
	5	.091	.989	.003	.011
	6	.107	.988	.003	.012
	8	.115	.987	.003	.013
	9	.131	.985	.004	.015
	10	.173	.980	.004	.020
	11	.191	.978	.004	.022
	12	.200	.977	.005	.023
	13	.209	.976	.005	.024
	14	.248	.971	.005	.029
	15	.299	.966	.006	.035

	16	.362	.959	.006	.042
	17	.418	.952	.007	.049
	18	.502	.943	.008	.059
	19	.541	.939	.008	.063
	20	.580	.934	.009	.068
	21	.608	.931	.009	.071
	22	.636	.928	.009	.074
	23	.726	.919	.010	.085
	24	.834	.907	.011	.097
	25	.914	.899	.011	.107
	26	1.078	.882	.012	.126
	27	1.233	.866	.013	.144
	28	1.359	.853	.014	.159
	29	1.620	.828	.015	.189
	30	2.061	.786	.016	.241
	31	2.516	.745	.018	.294
	32	2.773	.723	.019	.324
	33	3.075	.698	.019	.359
	34	3.303	.680	.020	.386
	35	3.639	.654	.021	.425
	36	4.080	.621	.021	.477
	37	4.545	.588	.022	.531
	38	4.809	.570	.023	.562
	39	5.262	.541	.024	.615
	40	5.660	.516	.024	.661
	41	6.115	.490	.025	.714
	42	6.420	.472	.025	.750
	43	6.862	.449	.026	.801
	44	7.657	.409	.028	.894
	45	7.939	.396	.028	.927
	46	8.301	.379	.029	.970
	47	9.915	.314	.032	1.158
	48	10.845	.282	.034	1.267
	50	11.746	.254	.036	1.372

	51	14.069	.193	.036	1.643
	52	15.550	.163	.034	1.816
	53	16.679	.143	.033	1.948
	54	17.838	.125	.032	2.083
	55	19.028	.108	.030	2.222
	56	21.047	.086	.027	2.458
	57	25.552	.051	.019	2.984
	58	28.403	.036	.016	3.317
	59	33.606	.020	.010	3.925
	61	36.099	.015	.009	4.216
	62	37.489	.013	.008	4.379
	63	42.046	.007	.005	4.911
	64	46.941	.004	.003	5.482
	65	50.611	.003	.002	5.911
	66	54.740	.002	.002	6.393
	67	59.297	.001	.001	6.926
	68	65.044	.001	.001	7.597
	69	68.116	.000	.000	7.956
	70	71.687	.000	.000	8.373
	72	77.982	.000	.000	9.108
	73	92.720	.000	.000	10.829
	76	101.539	.000	.000	11.859
New NOPF=5-10	0	.010	.999	.001	.001
	4	.039	.996	.002	.005
	5	.048	.994	.002	.006
	6	.058	.993	.003	.007
	7	.068	.992	.003	.008
	8	.089	.990	.003	.010
	9	.120	.986	.004	.014
	10	.215	.975	.005	.025
	11	.247	.972	.006	.029
	13	.302	.965	.006	.035
	14	.381	.956	.007	.044
	15	.497	.944	.008	.058
	16	.607	.932	.009	.071
	17	.734	.918	.010	.086

18	.876	.903	.011	.102
19	.995	.890	.011	.116
20	1.104	.879	.012	.129
21	1.276	.862	.013	.149
22	1.363	.853	.013	.159
23	1.452	.844	.013	.170
24	1.682	.822	.014	.196
25	1.958	.796	.015	.229
26	2.379	.757	.016	.278
27	2.591	.739	.017	.303
28	2.917	.711	.018	.341
29	3.439	.669	.018	.402
30	3.834	.639	.019	.448
31	4.278	.607	.020	.500
32	5.096	.551	.020	.595
33	5.610	.519	.020	.655
34	6.006	.496	.021	.701
35	6.510	.468	.021	.760
36	7.272	.428	.021	.849
37	7.893	.398	.021	.922
38	8.923	.353	.021	1.042
39	9.618	.325	.020	1.123
40	10.462	.295	.020	1.222
41	11.208	.270	.020	1.309
42	12.960	.220	.019	1.514
43	13.593	.204	.018	1.588
44	14.825	.177	.018	1.732
45	16.150	.152	.017	1.886
46	17.839	.124	.016	2.084
47	19.227	.106	.015	2.246
48	21.851	.078	.013	2.552
50	22.964	.068	.012	2.682
51	24.058	.060	.012	2.810
52	27.385	.041	.011	3.198
53	34.109	.019	.007	3.984
54	35.927	.015	.006	4.196
55	41.136	.008	.004	4.804
56	45.398	.005	.003	5.302
57	48.621	.003	.002	5.679

	58	50.293	.003	.002	5.874
	59	53.770	.002	.001	6.280
	61	65.811	.000	.000	7.686
	62	68.050	.000	.000	7.948
	63	72.770	.000	.000	8.499
	64	75.360	.000	.000	8.802
	65	87.402	.000	.000	10.208
	66	102.140	.000	.000	11.929
	67	110.959	.000	.000	12.959
	68	123.001	.000	.000	14.366
	70	139.599	.000	.000	16.305
	71	176.560	.000	.000	20.621
	73	195.561	.000	.000	22.841
New NOPF =11-20	1	.018	.998	.002	.002
	4	.037	.996	.003	.004
	5	.093	.989	.005	.011
	7	.113	.987	.005	.013
	9	.132	.985	.006	.015
	10	.172	.980	.007	.020
	11	.212	.976	.007	.025
	12	.233	.973	.008	.027
	13	.315	.964	.009	.037
	14	.486	.945	.011	.057
	15	.623	.930	.013	.073
	16	.722	.919	.014	.084
	17	.874	.903	.015	.102
	18	.976	.892	.016	.114
	19	1.055	.884	.016	.123
	20	1.272	.862	.018	.149
	21	1.504	.839	.019	.176
	22	1.565	.833	.020	.183
	23	1.695	.820	.020	.198
	24	1.988	.793	.022	.232
	25	2.194	.774	.023	.256
	26	2.632	.735	.025	.307
	27	2.963	.707	.026	.346
	28	3.391	.673	.027	.396
	29	4.085	.621	.029	.477
	30	4.589	.585	.029	.536

	31	5.061	.554	.030	.591
	32	5.882	.503	.031	.687
	33	6.722	.456	.031	.785
	34	7.791	.403	.031	.910
	35	8.699	.362	.031	1.016
	36	9.743	.320	.030	1.138
	37	10.471	.294	.030	1.223
	38	11.462	.262	.029	1.339
	39	12.387	.235	.028	1.447
	40	13.451	.208	.027	1.571
	41	14.471	.184	.026	1.690
	42	16.479	.146	.024	1.925
	43	17.684	.127	.023	2.065
	44	19.643	.101	.021	2.294
	45	21.439	.082	.019	2.504
	47	23.366	.065	.018	2.729
	48	27.851	.039	.014	3.253
	49	30.323	.029	.012	3.542
	50	33.349	.020	.011	3.895
	51	39.096	.010	.009	4.566
	53	45.391	.005	.006	5.301
	58	53.170	.002	.003	6.210
	59	72.170	.000	.000	8.429
	62	84.212	.000	.000	9.836
	63	98.950	.000	.000	11.557
	66	190.510	.000	.000	22.251
	16	.070	.992	.008	.008
	19	.141	.984	.011	.016
	20	.212	.976	.014	.025
	21	.657	.926	.023	.077
	22	1.050	.885	.028	.123
New NOPF =>20	23	1.131	.876	.029	.132
	24	1.381	.851	.032	.161
	25	2.000	.792	.036	.234
	26	2.188	.775	.037	.256
	27	2.679	.731	.040	.313
	28	3.430	.670	.042	.401

	29	4.270	.607	.044	.499
	30	4.663	.580	.045	.545
	31	4.938	.562	.045	.577
	32	6.856	.449	.044	.801
	33	7.445	.419	.044	.870
	34	9.308	.337	.042	1.087
	35	11.120	.273	.040	1.299
	36	11.844	.251	.040	1.383
	37	12.259	.239	.039	1.432
	38	13.170	.215	.039	1.538
	39	15.442	.165	.035	1.804
	40	16.926	.138	.033	1.977
	41	17.805	.125	.032	2.079
	42	18.973	.109	.031	2.216
	43	22.121	.076	.027	2.584
	53	24.932	.054	.026	2.912
	54	27.867	.039	.023	3.255
	55	41.121	.008	.008	4.803
	56	44.896	.005	.006	5.244
	57	53.163	.002	.003	6.209
	58	62.614	.001	.001	7.313
	59	79.942	.000	.000	9.337
	60	86.901	.000	.000	10.150
	65	94.680	.000	.000	11.058
	66	103.500	.000	.000	12.088
	71	113.681	.000	.000	13.277
	73	140.460	.000	.000	16.405
	74	186.240	.000	.000	21.752
	77	232.020	.000	.000	27.099